BLICATIONS DE L'INSTITUT DE STATISTIQUE
DE L'UNIVERSITÉ DE PARIS

MÉMOIRES ET CONFÉRENCES SUR LE CALCUL DES PROBABILITÉS,
LA STATISTIQUE THÉORIQUE ET APPLIQUÉE, L'ÉCONOMÉTRIE

HOMMAGE A GEORGES DARMOIS

Pierre THIONET

LA PERTE D'INFORMATION PAR SONDAGE

(Deuxième partie)

Daniel SCHWARTZ

LA MÉTHODE STATISTIQUE EN MEDECINE

VOL. IX FASCICULE 1 - 1960

PARIS
11, Rue Pierre-Curie

BLICATIONS DE L'INSTITUT DE STATISTIQUE DE L'UNIVERSITÉ DE PARIS

MÉMOIRES ET CONFÉRENCES SUR LE CALCUL DES PROBABILITÉS,
LA STATISTIQUE THÉORIQUE ET APPLIQUÉE, L'ÉCONOMÉTRIE

HOMMAGE A GEORGES DARMOIS

Pierre THIONET

LA PERTE D'INFORMATION PAR SONDAGE

(Deuxième partie)

Daniel SCHWARTZ

LA MÉTHODE STATISTIQUE EN MEDECINE

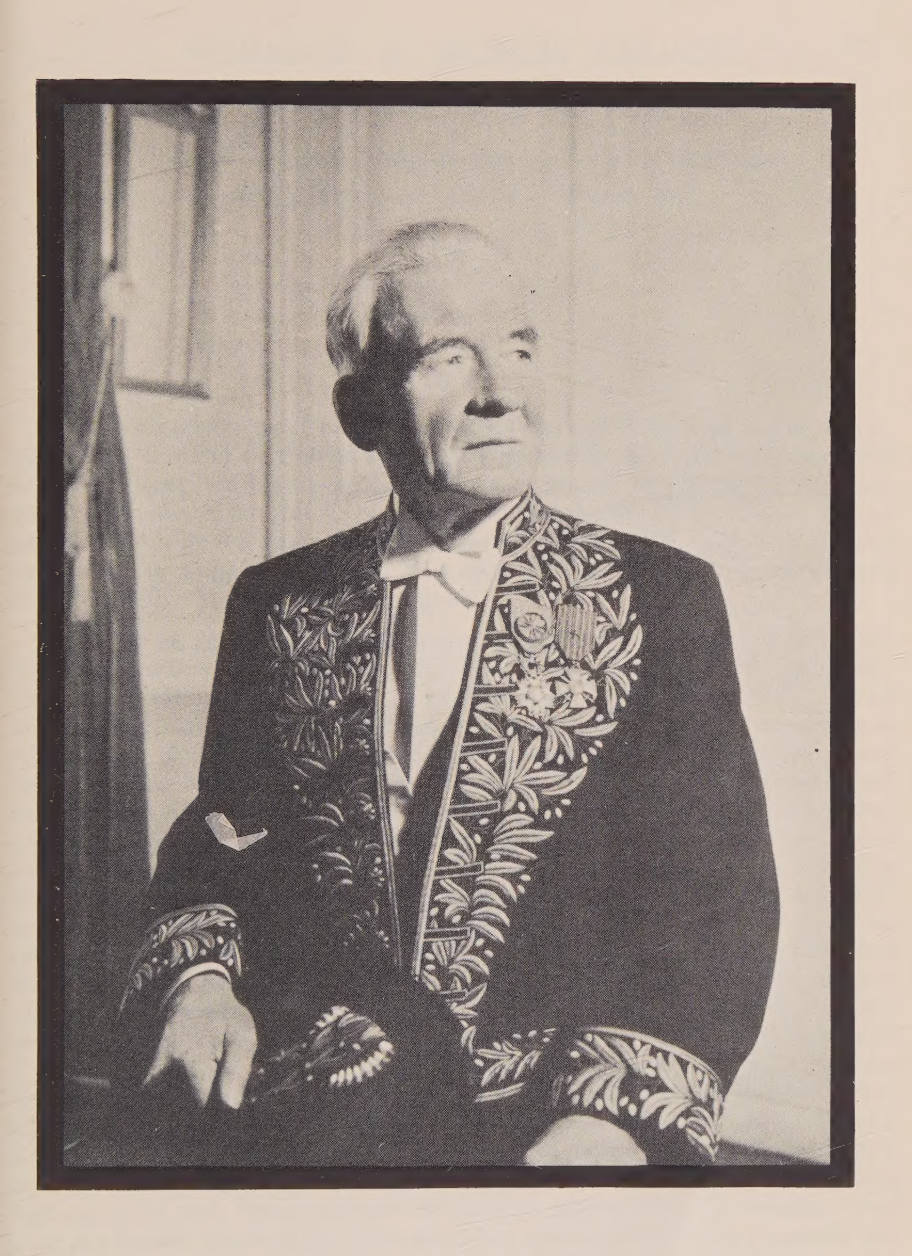
VOL. IX - FASCICULE 1 - 1960

PARIS

11, Rue Pierre-Curie

Toute la correspondance relative aux publications
doit être envoyée à l'adresse
INSTITUT DE STATISTIQUE DE L'UNIVERSITÉ DE PARIS
Institut Henri Poincaré - 11, Rue Pierre Curie - Paris (5e)

Les manuscrits doivent être envoyés à M. Daniel DUGUE
à l'adresse précédente.



HOMMAGE A GEORGES DARMOIS

(24 Juin 1888 - 3 Janvier 1960)

Prononcé au Conseil d'Administration de l'Institut de Statistique de l'Université de Paris, le 27 Janvier 1960

C'est en 1925 que Georges Darmois est entré dans le corps enseignant de l'Institut de Statistique de l'Université de Paris à la demande d'Emile Borel. Ce dernier récemment élu député de l'Aveyron le chargea d'assurer, à sa place, le cours de Calcul des Probabilités. Georges Darmois devenu plus tard titulaire de ce poste le conserva jusqu'à sa mort. Sa suppléance était au cours de cette année assurée par MM. Girault, Morlat et Indjoudjian.

A la retraite de Michel Huber en 1944 il fut nommé à la Direction des Etudes de l'Institut qu'il cumula avec le Secrétariat général. Dès la fin de la guerre 1914, le géomètre, le Spécialiste de la théorie de la relativité avait infléchi sa vie scientifique dans la direction du Calcul des Probabilités et plus spécialement de la Statistique. Deux Membres de notre conseil, MM. Bunle et Rueff ont, je crois, une part de responsabilité dans cette décision si heureuse pour la science française. En 1928 il publie un traité de statistique traduit plus tard en plusieurs langues dont le chinois et même l'anglais. Georges Darmois était très fier de le rappeler. On y retrouve le souci de clarté et d'élégance qui marque l'œuvre et la personnalité de l'auteur comme on les trouvera plus tard dans le livre "Statistique et Applications" publié chez Armand Colin. En dehors de ce travail d'exposition la Statistique mathématique est redevable à Georges Darmois de recherches très importantes. Il est un de ceux qui ont le plus étudié la notion d'exhaustivité d'une information, notion capitale puisque certains statisticiens comme Halphen et Savage m'ont avoué qu'ils ne concevaient la théorie de l'estimation que dans le cas où l'exhaustivité est possible. Georges Darmois a aussi perfectionné, et les beaux travaux de notre collègue Delaporte le qualifieraient beaucoup mieux que moi pour le dire, le modèle général mis au point par Spearman dit d'Analyse factorielle. Ce modèle créé pour les études psychologiques est en fait applicable à bien d'autres domaines et en particulier à l'économie mathématique. Georges Darmois enfin s'est intéressé brillamment à la théorie de la corrélation et de l'indépendance des variables aléatoires.

Ce grand savant fut aussi un grand administrateur, toute sa personnalité le qualifiait pour ce rôle, son ampleur de vues, sa générosité, son rayonnement, sa sérénité. Ce sont les termes dont se servent tous ceux qui ont eu à évoquer sa mémoire.

L'Institut de Statistique lui doit un développement qu'on osait à peine prévoir puisque les locaux qui nous sont affectés sont pleins à craquer parfois jusqu'à onze heures du soir.

Georges Darmois a créé à l'intérieur de l'Institut de Statistique un "Centre de formation des Ingénieurs et Cadres aux applications industrielles de la Statistique" et un Bureau universitaire de Recherche opérationnelle, des cours de formation statistique pour médecins.

Grâce à lui, je cite ici M. Geary, Conseiller Statistique des Nations Unies et ancien Vice Président de l'Institut International de Statistique, le retard que la France avait dans les applications de la Statistique a été plus que comblé.

Deux revues : La Revue de Statistique appliquée, dirigée par M. Morice et les Publications de l'Institut de Statistique qu'il avait bien voulu me confier lui doivent d'avoir vu le jour ainsi que l'Association des Anciens Elèves de l'Institut de Statistique, lien précieux entre l'enseignement et la vie quotidienne.

Ce fut enfin un soldat. La liaison étroite que comme Directeur de l'Institut de Statistique il a pu établir avec le Comité d'Action Scientifique de la Défense Nationale était certainement une de ses réalisations les plus chères. En 1938, ayant accompli les vingt huit ans de services militaires exigés de tous les français, il devait à moins de manifester expressément le désir contraire être rayé du Cadre des officiers de réserve et rendu définitivement à la vie civile. Il refusa de bénéficier de cette disposition de la loi, ce qui entrainait pour lui la charge de douze années de disponibilité militaire supplémentaires : la conjoncture internationale ne laissait aucun doute sur les risques que comportait une telle attitude. Au jour de la mobilisation générale, en septembre 1939, c'est donc comme volontaire que le Capitaine Georges Darmois rejoignait le célèbre 6ème groupe Autonome d'Artillerie, l'unité de formation des sections de repérage par le son, l'arme dans laquelle il avait avec autant de compétence que de courage apporté sa contribution à la victoire de 1918. Quelques semaines plus tard il était versé à la Mission scientifique franco-britannique organisée par M. Paul Montel à la demande de M. Dautry qui venait d'être nommé Ministre de l'Armement. Envoyé en mission à Londres le 15 Juin 1940 il demandait, mission accomplie, à regagner la France. Cette autorisation lui ayant été refusée il devait attendre le début de 1943 pour pouvoir s'installer

à Alger où sa science et son dévouement trouvaient évidemment à s'exercer facilement.

En particulier Georges Darmois éprouvait une légitime fierté d'avoir organisé les concours des grandes écoles en Afrique du Nord en 1943 et je m'honore de l'aide que j'ai pu lui apporter. De cette façon, était sauvegardé l'avenir de cette élite morale de la jeunesse française qui au prix de difficultés et de dangers sans nombre devait franchir les Pyrénées pour rejoindre l'armée. Il eut plus tard la grande joie de retrouver comme élèves d'anciens polytechniciens et d'anciens normaliens ainsi recrutés.

Le lundi 23 Novembre il s'asseyait pour la dernière fois dans son bureau de l'Institut Henri Poincaré. A son poste jusqu'au bout il reçut des visites, s'intéressant comme toujours au travail et à la carrière de ses élèves. C'était sa dernière sortie. Elle fut comme il l'eût sans doute souhaité consacrée à notre Institut. Inquiet de sa fatigue je le raccompagnai chez lui dans cette maison qu'il ne devait plus quitter où avec Madame Darmois, vers qui monte notre sympathie, il avait si souvent accueilli de son sourire cordial et amène tous ceux qui voulaient l'approcher. Et puis ce fut le lent déclin de ses forces sans que cette grande intélligence fût atteinte. Madame Soury, M. Bunle, Girault et moi-même qui venions le voir, espérions encore, habitués hélas à d'autres crises. C'est seulement le 1er Janvier au matin qu'à la suite d'un entretien téléphonique avec son ami de toujours M. Gustave Ribaud, j'ai compris qu'il fallait mettre son espérance ailleurs que dans sa guérison.

Ses ultimes préoccupations ont été scientifiques, historiques, et religieuses. Quand j'ai pris congé de lui ici-bas, le 18 Décembre, les dernières paroles qu'il m'ait adressées ont été pour me dire qu'au cours de la nuit précédente il avait essayé de percer à jour la pensée de Bayes, peu satisfait de ne connaître les idées du mathématicien britannique sur la Probabilité des causes que par le rapport de tiers, et qu'il avait réfléchi au problème de gouvernement que posait pour Saint-Louis la question de l'hommage.

La Bible était une de ses lectures favorites... et pendant que nous veillions sa dépouille, ma femme et moi nous évoquions le verset d'Esaïe qu'il aimait citer : "Sentinelle que dis-tu de la nuit ?"

Vous n'entendrez sans doute pas sans émotion la dernière phrase que sur la terre d'Afrique mon Maître consacrait à Newton en 1944 dans la Revue d'Alger. Georges Darmois concluait ainsi :"Et l'on doit à la France de dire que si Newton poursuit en l'autre monde des dialogues sur la mécanique céleste et l'Astronomie, c'est probablement avec des géomètres français qu'il s'entretient, car c'est en notre pays qu'on a surtout contribué à établir sa théorie et sa gloire".

L'institut de Statistique de l'Université de Paris conservera longtemps le souvenir de l'éminent humaniste scientifique qui l'a dirigé pendant quinze ans et qui a laissé un tel exemple à ses élèves.

M. D. DUGUE

Pierre THIONET

LA PERTE D'INFORMATION PAR SONDAGE

(Calcul des probabilités)

(Suite)

CHAPITRE V

LES SONDAGES A PLUSIEURS DEGRÉS

I - GENERALITES -

L'étude des sondages à deux degrés englobe comme cas limites :

- le sondage en grappe(1) si le 2ème degré disparaît;
- le sondage stratifié si le 1er degré disparaît.

Par sondage à plusieurs degrés on entend ici plus généralement une combinaison quelconque de sondages en degrés, avec stratifications, sous-stratifications, tirages en grappes, probabilités égales ou inégales, avec ou sans remise des boules tirées. Sont exclus les schémas de tirage avec probabilités inégales tels que les unités de sondage ne puissent jamais être prises qu'une seule fois(2).

Le découpage est supposé complet et invariable. Enfin les divers tirages au sort sont supposés indépendants entre eux. A chaque tirage, le nombre d'unités de sondage à tirer est prédéterminé, indépendant des tirages précédents, que ce nombre soit fixe ou en proportion fixe de certains paramètres de la population et du sondage. De plus chaque tirage est indépendant des tirages voisins : dans les autres strates et les autres unités de sondage de même rang.

Toutes ces hypothèses sont habituelles en technique de sondage; par exemple établir des compensations entre strates (Goodman et Kish) n'est pas d'usage courant.

----

(1) Echantillon en grappes - Mêmes formules que pour le sondage à 1 degré, en désignant par σ l'écart-type entre les grappes et non entre les unités.

(2) Car il en résulte des complications que les travaux les plus récents commencent seulement à surmonter (voir T. 5, ch. IV).

II - ECHANTILLON STRATIFIE -

Considérons une population découpée en strates numérotées 1, 2, ...h... Adoptons les notations suivantes :

Strate h	Population
effectif : ν_h	$\nu = \Sigma_h \nu_h$
moyenne des x_{hi} : $\bar{x}_h$	$\bar{x} = \Sigma_h \bar{x}_h / \nu$
écart-type des x_{hi} : σ_h	estimateur sans biais de $\bar{x}$:
effectif de l'échantillon : n_h	
moyenne des x_{hi} échantillon : $\bar{X}_h$	$\bar{X} = \Sigma_h \frac{\nu_h}{\nu} \bar{X}_h$

La variance est

$$\mathcal{V}\bar{X} = \sum_n \left(\frac{\nu}{\nu}\right)^2 \mathcal{V}(\bar{X}_h)$$

c'est-à-dire

$$= \sum_n \left(\frac{\nu_h}{\nu}\right)^2 p(\nu_h, n_h)$$

Si l'on envisage, toutes choses égales d'ailleurs, un second échantillon d'effectifs n'_h tels qu'on ait (quel que soit h)

$$n'_h \leqslant n_h$$

on aura :

$$\mathcal{V}\bar{X}' - \mathcal{V}\bar{X} = \sum_h \left(\frac{\nu_h}{\nu}\right)^2 [p(\nu_h, n'_h) - p(\nu_h, n_h)]$$
$$= \sum_h \left(\frac{\nu_h}{\nu}\right)^2 \mathcal{E}P(n_h, n'_h)$$

Et si l'on pose, symboliquement cette fois :

$$\left.\begin{aligned} \mathcal{V}\bar{X} &= p(\nu, n) \\ \mathcal{V}\bar{X}' &= p(\nu, n') \end{aligned}\right\} \quad \sum_h \left(\frac{\nu_h}{\nu}\right)^2 p(n_h, n'_h) = P(n, n')$$

on aura bien la relation :

$$\boxed{p(\nu, n') = p(\nu, n) + \mathcal{E}P(n, n')}$$

La variance $\mathcal{V}\bar{X}$ est une perte d'information.

Topologie du sondage stratifié.

Figurons tous les estimateurs $\bar{X}$ (ci-dessus) par des points sur un demi-axe, points rangés par ordre de variance croissante; c'est la représentation la plus simple.

Par exemple, on peut porter, à partir de l'origine U, la longueur $\mathcal{V}\bar{X} = (U\bar{X})$; la distance $(\bar{X}\bar{X}')$ de deux points de l'axe est égale à la combinaison des suppléments de perte d'information par strate, tandis que $(\mathcal{V}\bar{X})$ est la combinaison des pertes d'information par strate (combinaisons linéaires à coefficients constants).

Toutefois l'ordre des $\bar{X}$ sur leur axe dépend des grandeurs respectives des paramètres σ_h et $\sigma_{h'}$ de chaque strate. Pour obtenir une figure invariante, on représentera les sondages stratifiés par le grillage (n_h) : à chaque sommet de ce grillage correspond l'estimateur sans biais $\bar{X}$ du plan de sondage correspondant.

Remarque 1.

Si les tirages sont avec remise, les "barreaux" du grillage sont en nombre infini : ils s'accumulent, leurs limites correspondant à la connaissance complète d'une certaine strate.

Remarque 2.

Les "barreaux" du grillage (sur toute leur longueur) sont axes d'information : l'échantillon extrait de toutes les strates, sauf une, restant fixe, la variation de l'effectif de cette seule strate correspond à des tirages dans une urne unique.

Le seul terme de la variance $\mathcal{V}\bar{X}$, qui correspond à cette strate, varie; c'est $\mathcal{V}\bar{X}_h$ (au facteur $(\nu_h/\nu)^2$ près); $\mathcal{V}\bar{X}$ vérifie donc bien la condition d'alignement.

Représentation métrique multidimensionnelle. Usage d'une dimension par strate.

Dans l'espace euclidien (E) à L dimensions, (par exemple à 3 dimensions pour 3 strates), on conviendra que le point $\bar{X}$ de coordonnées ξ_h

$$\xi_h = \left(\frac{\nu_h}{\nu}\right)^2 \mathcal{V}\bar{X}_h$$

représente l'estimateur $\bar{X}$, l'origine des coordonnées représentant $\bar{x}$

D'où :

$$\mathcal{V}\bar{X} = \Sigma_h \ \xi_h$$

Plus généralement, le point $\bar{X}$ représente le plan de sondage (n_h)

Lorsqu'on donne aux divers n_h toutes les valeurs permises 1, 2,

... ν_h , l'ensemble des points $\bar{X}$ constitue les "nœuds" ou sommets d'un grillage.

Et lorsqu'on fixe les valeurs de tous les n_h sauf un, le sous-ensemble de points s'aligne sur une demi-droite constituant un "barreau" dudit grillage.

On ne nuit pas à la généralité en se limitant ici à 3 strates.

a) Soit $\bar{X}$ et $\bar{X}'$ deux estimateurs représentés par deux points du même "barreau", - disons que :

$$n_1 > n'_1 ; \qquad n_2 = n'_2 ; \qquad n_3 = n'_3 .$$

On aura :

$$\mathcal{V}\bar{X}' - \mathcal{V}\bar{X} = \mathcal{E}(\bar{X}' - \bar{X})^2 = \left(\frac{\nu_1}{\nu}\right)^2 \frac{\nu_1 \ \sigma_1^2}{\nu_1 - 1}\left(\frac{1}{n'_1} - \frac{1}{n_1}\right)$$

b) Supposons à présent $n_1 > n'_1$; $n_2 > n'_2$; $n_3 = n'_3$. On peut définir un point intermédiaire $\bar{X}''$ (n_1, n'_2, n_3) de façon que :

$$\mathcal{V}\bar{X}' - \mathcal{V}\bar{X}'' = \mathcal{E}(\bar{X}' - \bar{X}'')^2 = \left(\frac{\nu_1}{\nu}\right)^2 \frac{\nu_1 \ \sigma_1^2}{\nu_1 - 1}\left(\frac{1}{n'_1} - \frac{1}{n_1}\right)$$

$$\mathcal{V}\bar{X}'' - \mathcal{V}\bar{X} = \mathcal{E}(\bar{X}'' - \bar{X})^2 = \left(\frac{\nu_2}{\nu}\right) \frac{\nu_2 \ \sigma_2^2}{\nu - 1}\left(\frac{1}{n'_2} - \frac{1}{n_2}\right)$$

$$\mathcal{V}\bar{X}' - \mathcal{V}\bar{X} = \mathcal{E}(\bar{X}' - \bar{X}'')^2 + \mathcal{E}(\bar{X}'' - \bar{X})^2 = \mathcal{E}(\bar{X}' - \bar{X})^2$$

Ici les segments $\bar{X}\bar{X}''$ et $\bar{X}''\bar{X}'$ sont parallèles aux axes; dans le cas général, c'est tout une ligne polygonale $\bar{X}\bar{X}''$, $\bar{X}''\bar{X}''$, ... , $\bar{X}''\bar{X}'$, dont chaque segment est parallèle à un axe de coordonnées, qu'il y aurait lieu de considérer.

Une confusion à éviter.

On évitera de confondre l'espace (E) considéré ci-dessus avec l'espace (A) dont on fait aussi usage à propos de sondage. Tout aléatoire y est figuré par un vecteur. Deux aléatoires indépendants sont figurés par deux vecteurs orthogonaux de (A), l'espérance mathématique du carré de la différence entre deux vecteurs correspond au carré de la distance dans (A).

Par exemple, $\bar{X}$ et $\bar{X}'$ étant définis comme ci-dessus, on a :

$$\mathcal{E}(\bar{X}' - \bar{X})^2 = \mathcal{E}(\bar{X}' - \bar{X})^2 + \mathcal{E}(\bar{X} - \bar{x})^2$$

$$\mathcal{E}(\overline{X}' - \overline{X}) \cdot (\overline{X} - \overline{x}) = 0$$

relation qui correspond à l'orthogonalité des vecteurs $\overline{x}\overline{X}$ et $\overline{X}\overline{X}'$ dans l'espace (A) évoqué, et non dans l'espace euclidien (E).

Bien au contraire, on a vu dans l'exemple (a) que $\overline{X}\overline{X}'$ pouvait être parallèle à un axe, alors que $\overline{x}\overline{X}$ n'était pas forcément parallèle à l'autre axe, et d'ailleurs sans qu'on ait exigé que les axes de coordonnées de (E) soient rectangulaires.

On peut en particulier en déduire le paradoxe suivant. Posons :

$$X^* = \lambda \overline{X} + (1 - \lambda)\overline{X}'$$

estimateur sans biais de $\overline{x}$ au même titre que $\overline{X}$ et $\overline{X}'$; dans l'espace (A) il ne fait aucun doute que le point représentant cette combinaison linéaire de $\overline{X}$ et $\overline{X}'$ est sur le segment joignant les points $\overline{X}$ et $\overline{X}'$. On pourrait donc croire que, dans l'espace euclidien (E) il en serait de même. Ce serait inexact.

Le calcul montre immédiatement que :

$$\mathcal{E}(X^* - \overline{X})^2 = (1 - \lambda)^2 \mathcal{E}(\overline{X}' - \overline{X})^2$$

$$\mathcal{E}(X^* - \overline{X}')^2 = \lambda^2 \mathcal{E}(\overline{X} - \overline{X}')^2$$

et il est donc impossible qu'on ait :

$$\text{distance } \overline{X}\overline{X}' = \text{distance } \overline{X}X^* + \text{distance } X^*\overline{X}'$$

lorsque $\overline{X}$ et $\overline{X}'$ sont (comme en a) sur un barreau du grillage.

Sondages stratifiés particuliers.

a) Sondages représentatifs.

De l'ensemble des points $\overline{X}$ ou (n_h) on va extraire les points $(f\,\nu_h)$ sous-ensemble à un seul paramètre f; dans l'espace (E), les ξ_h sont fonctions linéaires de $1/f$ et les points sont alignés.

$$\xi_h = a_h\left(\frac{1}{f} - 1\right)$$

Lorsque $f\nu_h$ n'est pas entier, on peut convenir de rejeter ou non les points. On retiendra que le sondage "représentatif" est représenté par une division sur un axe, homothétique de celle du sondage simple.

b) Sondage "optimum" au sens de Neyman et Yates.

On sait qu'en adoptant des n_h proportionnels aux $\nu_h\ \sigma_h/\sqrt{C_h}$, on rend la variance $\mathcal{V}\bar{X}$ minimum à coût constant $\Sigma\ C_h n_h = C$.

Les points (n_h) correspondants (après arrondissement des n_h) se trouvent sensiblement eux aussi sur un axe d'information :

$$\xi_h = b_h - \frac{1}{C} b'_h$$

On sait effectivement que $\mathcal{V}\bar{X}$ a pour expression :

sondage représentatif : $\sum \left(\frac{\nu_h}{\nu}\right)^2 \frac{\sigma_h^2}{\nu_h - 1}\left(\frac{1}{f} - 1\right)$;

sondage "optimum" : $\sum \left(\frac{\nu_h}{\nu}\right) \frac{\sigma_h^2}{\nu_h - 1} - \frac{1}{C}\left[\sum \frac{\nu_h}{\nu}\sigma_h\sqrt{\frac{\nu_h\ C_h}{\nu_h - 1}}\right]$

III - SONDAGE A 2 DEGRES A 2 PARAMETRES, AVEC TIRAGES EQUIPROBABLES -

On a indiqué déjà (Ch. III, § II) que le sondage à 2 degrés et 2 paramètres $(m, \bar{n})$ (avec $\nu_i = \bar{\nu}$) correspondait au treillis le plus simple : le grillage $(m, \bar{n})$ dont les sommets sont repérés par les entiers $(m, \bar{n})$, $1 \leqslant m \leqslant \mu$, $1 \leqslant \bar{n} \leqslant \bar{\nu}$.

Le point $(\mu, \bar{\nu})$ représente la connaissance parfaite; tandis que le point $(m, \bar{n})$ représente les sondages de taille $(m, \bar{n})$.

Posons : $\bar{X} = S_i S_j x_{ij}/m\,\bar{n}$; $\quad \bar{\bar{x}} = \Sigma_i \Sigma'_j x_{ij}/\mu\,\bar{\nu}$

$\bar{x} = S_i \bar{x}_i/m$, avec $\bar{x}_i = \Sigma_j x_{ij}/\bar{\nu}$

$\bar{\bar{X}} = \Sigma_i \bar{X}_i/\mu$, avec $\bar{X}_i = S_j x_{ij}/\bar{n}$

Attachons $\bar{X}$ à $(m, \bar{n})$ et $\bar{\bar{x}}$ à $(\mu, \bar{\nu})$, $\bar{x}$ à $(m, \bar{\nu})$ et $\bar{\bar{X}}$ à $(\mu, \bar{n})$.

Nous avons montré ailleurs (T. 4) que, pour tous les sondages à deux degrés (indépendants l'un de l'autre), on avait :

$$\mathcal{V}\bar{X} = \mathcal{E}(\bar{X} - \bar{\bar{x}})^2 = \mathcal{E}(\bar{X} - \bar{x})^2 + \mathcal{E}(\bar{x} - \bar{\bar{x}})^2$$

$$= \mathcal{E}(\bar{X} - \bar{\bar{X}})^2 + \mathcal{E}(\bar{\bar{X}} - \bar{\bar{x}})^2$$

c'est-à-dire :

$$\mathcal{E}(\overline{X} - \overline{x}) \,.\, (\overline{x} - \overline{\overline{x}}) = 0$$

$$\mathcal{E}(\overline{X} - \overline{\overline{X}}) \,.\, (\overline{\overline{X}} - \overline{\overline{x}}) = 0$$

A la première décomposition de $\mathcal{V}\,\overline{X}$ correspond une première famille d'axes :

$$\mathcal{E}(\overline{x} - \overline{\overline{x}})^2 \equiv \frac{\sigma_o^2}{m} \; \frac{\mu - m}{\mu - 1} = \text{constante}$$

donc

$$m = \text{constante}; \quad \overline{n} \text{ variable (verticales)}$$

avec

$$\mathcal{V}\,\overline{X} - \mathcal{E}(\overline{x} - \overline{\overline{x}})^2 \equiv \frac{1}{m\mu} \sum_i \frac{\overline{\nu}\,\sigma_i^2}{\overline{\nu} - 1} \left(\frac{1}{\overline{n}} - \frac{1}{\overline{\nu}}\right)$$

A la deuxième décomposition de $\mathcal{V}\overline{X}$ correspond une deuxième famille d'axes

$$\left\{ \begin{array}{l} \mathcal{E}(\overline{X} - \overline{\overline{x}})^2 \equiv \dfrac{1}{\mu^2} \; \dfrac{\Sigma\, \sigma_i^2}{\overline{n}} \; \dfrac{\overline{\nu} - \overline{n}}{\overline{\nu} - 1} = \text{constante} \\[2ex] \overline{n} = \text{constante; } m \text{ variable (horizontales)} \end{array} \right.$$

avec $$\mathcal{V}\,\overline{X} - \mathcal{E}(\overline{\overline{X}} - \overline{\overline{x}})^2 \equiv \frac{\mu\,\sigma_o^2}{\mu - 1} + \frac{1}{\mu} \sum \frac{\overline{\nu}\;\sigma_o^2}{\overline{\nu} - 1} \left(\frac{1}{\overline{n}} - \frac{1}{\overline{\nu}}\right)\left(\frac{1}{m} - \frac{1}{\mu}\right)$$

La formule de récurrence est valable pour $\mathcal{V}\,\overline{X}$ à condition de compter le trajet $\overline{X}\,\overline{X}'$ le long d'axes d'information et jamais directement entre deux points qu'un axe ne relie pas. Entre les points $(m, \overline{n})$ et $(m', \overline{n})$, $m > m'$, $\overline{n} > \overline{n}'$ on doit suivre un trajet en ligne brisée empruntant les côtés du grillage.

Remarque.

On a sur les premiers axes

$$\vartheta(\nu) - \vartheta(\pi) = \left[\frac{\overline{\nu} - 1}{\overline{\nu}} - \frac{\overline{n} - 1}{\overline{n}}\right] = \frac{1}{\overline{n}} - \frac{1}{\overline{\nu}}$$

et sur les seconds

$$\vartheta(\upsilon) - \vartheta(\pi) = \left[\frac{\mu - 1}{\mu} - \frac{m - 1}{m}\right] = \frac{1}{m} - \frac{1}{\mu}$$

Représentation métrique

Dans l'espace euclidien, considérons le point de coordonnées (α, β, γ)

$$\alpha = \mathcal{E}(\bar{x} - \bar{\bar{x}})^2; \qquad \beta = \mathcal{E}(\bar{\bar{X}} - \bar{\bar{x}})^2; \qquad \gamma = \left(\frac{\mu}{m} - 1\right)\beta$$

On a par ailleurs :

$$\alpha = \frac{\sigma_o^2}{\mu - 1}\left(\frac{\mu}{m} - 1\right)$$

donc :

$$\frac{\sigma_o^2}{\mu - 1}\,\gamma = \alpha\,\beta$$

équation d'un paraboloïde hyperbolique P.H. dont les deux familles de génératrices correspondent respectivement à α = constante et β = constante; c'est-à-dire se projettent suivant le grillage (m, $\bar{n}$) sur ($\alpha\,\beta$).

Double décomposition de $\mathcal{V}\bar{X}$.

$$\mathcal{V}\bar{X} = \alpha + (\beta + \gamma) = \mathcal{E}(\bar{x} - \bar{\bar{x}})^2 + \frac{\mu}{m}\left[\frac{1}{\mu^2}\,\frac{\Sigma\sigma_i^2}{\bar{\bar{n}}}\,\frac{\bar{\nu} - \bar{n}}{\bar{\nu} - 1}\right]$$

$$= \mathcal{E}(\bar{x} - \bar{\bar{x}})^2 + \mathcal{E}(\bar{X} - \bar{x})^2$$

$$\mathcal{V}\bar{X} = (\alpha + \gamma) + \beta = \mathcal{E}(\bar{X} - \bar{\bar{X}})^2 + \mathcal{E}(\bar{\bar{X}} - \bar{\bar{x}})^2$$

car $\mathcal{E}(\bar{X} - \bar{\bar{X}})^2 = \mathcal{E}(\bar{x} - \bar{\bar{x}})^2 + \mathcal{E}[(\bar{X} - \bar{x})\quad(\bar{\bar{X}} - \bar{\bar{x}})]^2 = \alpha + \gamma$

En résumé les plans de sondage à deux degrés, dans le cas où on a :

$$n_i = \bar{n}, \qquad \nu_i = \bar{\nu} \quad ,$$

sont représentés par les nœuds d'un treillis dessiné sur un certain paraboloïde hyperbolique (P.H.), treillis dont la projection sur le plan des (α, β) est un grillage.

Les côtés des mailles du treillis sont tous portés par des génératrices (de l'un ou l'autre système) de P.H.

La perte d'information est la somme ($\alpha + \beta + \gamma$) des trois coordonnées des points figuratifs.

Les points situés sur des plans parallèles, d'équation :

$$\alpha + \beta + \gamma = \text{constante}$$

représentent des sondages donnant d'égales pertes d'information.

Autre métrique.

Représentons le même phénomène avec des coordonnées α', β', γ' égales à $\sqrt{\alpha}\sqrt{\beta}\sqrt{\gamma}$. C'est même l'idée la plus naturelle, quand on a l'habitude de considérer $\mathcal{E}(\overline{X} - \overline{\overline{x}})^2$, etc. comme les carrés de certaines distances, et les relations

$$\mathcal{E}(\overline{X} - \overline{\overline{x}})^2 = \mathcal{E}(\overline{X} - \overline{\overline{X}})^2 + \mathcal{E}(\overline{\overline{X}} - \overline{\overline{x}})^2$$

$$= \mathcal{E}(\overline{X} - \overline{x})^2 + \mathcal{E}(\overline{x} - \overline{x})^2$$

comme deux expressions du théorème de Pythagore.

Alors la perte d'information est : $\alpha'^2 + \beta'^2 + \gamma'^2$ c'est-à-dire le carré de la distance de $\overline{\overline{x}}$ à $\overline{X}$.

Il est à noter que (P.H.) se transforme en un autre paraboloïde hyperbolique P.H' d'équation :

$$\frac{\sigma_0}{\sqrt{\mu - 1}} \gamma' = \alpha' \beta'$$

toutes les propriétés relatives aux treillis, grillage et génératrices rectilignes se conservent. Mais les lieux des points d'égales pertes d'information sont les sphères de centre $\overline{\overline{x}}$.

Remarque.

On évitera de confondre P.H. ou P.H' et l'espace (A) des sondages où, $\overline{X}_1$ et $\overline{X}_2$ étant représentés par deux points d'une même génératrice de P.H. ou P.H', on aurait :

$$(\overline{\overline{x}}\,\overline{X}_2)^2 = (\overline{\overline{x}}\,\overline{X}_1)^2 + (\overline{X}_1\,\overline{X}_2)^2$$

Il serait absolument vain de vouloir représenter cet espace dans l'espace à trois dimensions ; chaque génératrice du P.H. représente elle-même un espace euclidien à $\overline{\nu}$ dimensions.

IV - LES SONDAGES A PLUSIEURS DEGRES EN GENERAL -

1/ - Tirages équiprobables : Cas de deux degrés de sondage.

La double décomposition de la variance a été démontrée dans le cas général, à la seule condition d'écrire :

$$\mathscr{E}(\overline{X} - \overline{x})^2 = \frac{1}{m\mu} \sum_i \frac{\sigma_i^2}{n_i} \frac{\nu_i - n_i}{\nu_i - 1}$$

et en supposant exhaustifs les tirages élémentaires.

Au lieu de deux paramètres $(m, \overline{n})$, il y a $(\mu + 1)$ paramètres (m, n_i), car on suppose les n_i définis pour toutes les unités du 2ème degré, tirées ou non.

Si m ne change pas, la variation des n_i a le même effet que sur un sondage stratifié, de sorte que $\mathcal{V}\overline{X}$ se comporte comme une perte d'information au sens du chapitre I, représentable sur un axe. A chaque m correspond un tel axe, et leur ensemble peut être figuré par des barreaux parallèles empruntés à un grillage. Pourtant chaque axe symbolise un grillage à μ dimensions comme pour un sondage stratifié.

Si l'on fait varier m tout seul sans toucher aux n_i, la seconde décomposition de la variance entre en jeu; et il existe une seconde famille d'axes (n_i) représentée par les seconds barreaux du grillage, croisés avec les premiers. Ces axes ne sont d'ailleurs pas ordonnés, de même que l'ensemble des points (n_i).

Pour passer d'un plan de sondage à un autre (d'effectif plus réduit) on se déplacera sur une génératrice d'un système, puis sur une génératrice de l'autre système; et on aura encore le droit de dire que la variance est une perte d'information.

Représentation euclidienne métrique.

α n'est pas modifié et β ne l'est guère :

$$\beta = \frac{1}{\mu} \sum_i \frac{\sigma_i^2}{n_i} \frac{\nu_i - n_i}{\nu_i - 1}$$

On a encore :

$$\gamma = \frac{\mu - m}{m} \beta$$

Conclusion.

On retrouve le même P.H. qu'en II, et la même famille (dénombrable) de génératrices m = constante; en revanche sur le deuxième faisceau de génératrices, la substitution des n_i à n augmente très notablement le nombre de génératrices du P.H. réellement utilisées comme axes d'information.

2/ - Extension à un nombre quelconque de degrés de sondage.

On sait que la variance d'un sondage à trois degrés et plus se décompose (et de plusieurs façons) comme celle du sondage à deux degrés. Il est possible de représenter les sondages à d degrés par les nœuds d'un grillage généralisé à d dimensions.

Sur un barreau quelconque du grillage, $\mathcal{V}\,\overline{X}$ satisfait à la condition de récurrence. Pour comparer deux sondages quelconques, il faudra se déplacer de proche en proche le long des barreaux. Si les effectifs de l'un sont tous au moins égaux à ceux de l'autre, on peut alors écrire la condition de récurrence.

3/ - Cas où les tirages élémentaires sont bernoulliens.

On sait que la double décomposition de la variance du sondage à deux degrés est encore valable avec des tirages bernoulliens. La représentation par P.H. est encore valable :

a) en remplaçant $(\mu - 1)$ par μ dans l'équation du P.H. ;

b) en faisant intervenir une infinité dénombrable de génératrices du P.H. de chaque famille $(1 \leqslant \overline{n},\ 1 \leqslant m)$ et non plus un nombre fini de génératrices.

Les grillages comprennent une infinité dénombrable de barreaux des deux systèmes.

On peut accroître le nombre de degrés de sondage, - ou employer les sondages bernoulliens pour certains degrés et les sondages exhaustifs pour d'autres degrés.

4/ - Cas où il existe des strates et des sous-strates.

A condition d'adapter les axes utilisés, rien n'empêche de compliquer le plan de sondage avec des strates (au 1er degré), des sous-strates (au 2ème degré), etc.

5/ - Cas où les tirages ne sont plus équiprobables.

Limitons-nous au cas le plus courant où les probabilités de tirage au 1er degré sont proportionnelles à la taille des unités primaires de sondage et où une seule unité primaire est tirée; au second degré, les tirages sont supposés équiprobables.

On sait que la double décomposition de la variance est valable avec des probabilités de tirage inégales, à condition de modifier légèrement les définitions de $\overline{\overline{X}}$ et $\overline{\overline{x}}$ (moyennes pondérées pour $\overline{\overline{X}}$ et $\overline{\overline{x}}$, simples pour $\overline{X}$ et $\overline{x}$).

On peut donc étendre la théorie du grillage et aussi celle du paraboloïde hyperbolique.

$\overline{X}$ suit la formule de récurrence le long des axes d'information, et cette formule s'étend à un trajet en ligne brisée empruntant les axes ou génératrices; ceci suppose que la taille de l'échantillon s'est réduite (c'est-à-dire qu'aucun des effectifs n'augmente quand certains diminuent).

V - APPLICATION A L'ETUDE DU SONDAGE A 3 DEGRES (ET A 3 PARAMETRES) -

1/ - Décomposition de la variance en 7 composantes.

On abandonnera les notations précédentes (Réf.T.4) pour alléger. Si on désigne par X_{III} ou III l'estimateur sans biais de X_{000} ou 000, la variance de cet estimateur s'écrit :

$$V = E(X_{III} - X_{000})^2 = E(III - 000)^2$$

$$= E(III - 0II)^2 + E(0II - 00I)^2 + E(00I - 000)^2$$

etc.

Elle peut donc prendre six expressions différentes, qu'il est commode de représenter comme suit (avec 12 composantes) :

$a + B + \gamma$	$a + C' + \beta$
$b + C + \alpha$	$b + A' + \gamma$
$c + A + \beta$	$c + B' + \alpha$

Toutefois on a :

$$a + B = b + A' = a + b + (ab) = V - \gamma$$

$$b + C = c + B' = b + c + (bc) = V - \alpha$$

$$c + A = a + C' = c + a + (ca) = V - \beta$$

Ainsi la variance V s'exprime à l'aide de 9 composantes seulement : a, b, c, (ab), (bc), (ca), α, β, γ.

De même, il est clair que :

$$V - a = B + \gamma = C' + \beta = B + C' + (BC')$$

$$V - b = C + \alpha = A' + \gamma = C + A' + (CA')$$

$$V - c = A + \beta = B' + \alpha = A + B' + (AB')$$

En combinant ces résultats, il vient :

$$V = a + [b + (ab)] + [c + (ca)] + (BC')$$

$$= a + b + c + (ab) + (ca) + (bc) + (BC') - (bc)$$

et deux formules analogues; d'où il suit que :

$$(BC') - (bc) = (CA') - (ca) = (AB') - (ab) = z$$

et trois formules semblables, d'où trois expressions symétriques :

$$(B'C) - (\beta\alpha) = (C'A) - (\gamma\alpha) = (A'B) - (\alpha\beta)$$

On posera :

$$z = (abc)$$

en remarquant qu'avec le sondage à 2 degrés, les 3 composants de la variance V sont représentés par le symbole :

$$a + b + (ab) = (1 + a) \cdot (1 + b) - 1$$

et qu'avec le sondage à 3 degrés, le symbole à employer est, cette fois :

$$(1 + a) \cdot (1 + b) \cdot (1 + c) - 1 = a + b + c + (ab) + (bc) + (ca) + (abc)$$

2/ - Théorème.

Les 7 composantes précédentes de V sont des variances (et par conséquent sont positives).

En effet :

(1) (ab) (bc) (ca) sont par définition des <u>covariances</u> :

011 α 111
B' A β
001 101 γ
c C C'
b 010 A' 110
B
000 a 100

$$(ab) = B - b = A' - a$$

$$= E\ (110 - 100)^2 - E(010 - 000)^2$$

$$= E\ (110 - 010)^2 - E(100 - 000)^2$$

$$= E\ [(110+000-100-010)(110+010-100-000)]$$

$$= E\ [(110+000-100-010)(110+100-010-000)]$$

(2) Mais la théorie du sondage à 2 degrés[1] établit que (ab) est aussi une variance

$$(ab) = E(110 - 100 - 010 + 000)^2$$

Ceci résulte du choix d'estimateurs sans biais 110, 100 et 010,

qui entraîne l'orthogonalité de (110 - 100) et (100 - 000)

et celle de (110 - 010) et (010 - 000)
(comme sur le cube de la figure ci-dessus)

$$E\ [(110 - 100).(100 - 000)] = E\ [(110 - 010).(010 - 000)] = 0$$

avec

$$E\ [(100 - 000).(010 - 000)] = E\ [(110 - 100).(110 - 010)] = 0$$

D'où

$$E\ [(110-100).(010-000)] = E\ [(110-010+010-000+000-100)(010-000)]$$

$$= 0 + E\ (010 - 000)^2 + 0$$

D'où :

$$E(110-100-010+000)^2 = E(110-100)^2 + E(010-000)^2 - 2E\ [(110-100)(010-000)]$$

$$= E(110-100)^2 + E(010-000)^2 - 2E(010-000)^2$$

$$= E(110-100)^2 - E(010-000)^2$$

c. q. f. d.

(1) Cf. Annexe de l'Etude théorique n°6 de l'I. N. S. E. E. (Réf. T. 4).

(3) Or il est clair que (abc) joue vis-à-vis de (ab) (bc) et (ac) le même rôle que (ab) vis-à-vis de a et b. Cette différence de deux variances est encore une variance du fait de l'orthogonalité (c'est-à-dire de l'absence de biais des estimateurs).

C.Q.F.D.

Remarque.

Par exemple, dans les cas d'ordre r = 1, avec des tirages sans remise équiprobables, on sait que (ab) est de la forme :

$$(ab) = B - b = \frac{H}{\ell} - \frac{H}{\lambda} = H\left(\frac{I}{\ell} - \frac{I}{\lambda}\right) = M$$

De même :

$$(abc) = (AB') - (ab) = \frac{M}{n} - \frac{M}{\nu} = M\left(\frac{1}{n} - \frac{1}{\nu}\right)$$

Si l'on se souvient que :

$$b = \frac{H}{\lambda} = \frac{K}{\lambda}\left(\frac{1}{m} - \frac{1}{\mu}\right)$$

On arrive finalement à :

$$(abc) = K\left(\frac{1}{\ell} - \frac{1}{\lambda}\right).\left(\frac{1}{m} - \frac{1}{\mu}\right).\left(\frac{1}{n} - \frac{1}{\nu}\right)$$

3/ - Interprétation géométrique.

a) Dans l'espace à 3 dimensions, un parallélipipède de sommets (000), (001), ... (111), dont les 12 arêtes s'appelent a b c, A B C, A' B' C', α β γ, donne du plan de sondage une représentation topologique valable. La famille des plans de sondage est représentée par un grillage à 3 dimensions.

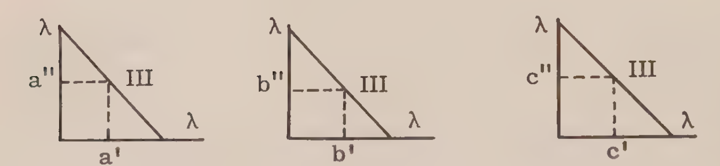
b) La représentation métrique nécessite au contraire un espace euclidien dont le nombre de dimensions n'est pas évident. On pourrait en prendre 9.

3 pour a b c;

3 pour (bc) (ca) (ab) ou a' b' c' ;

3 pour α - a, β - b, γ - c ou a'' b'' c''

c) Mais considérons les trois plans de coordonnées a' a'', b' b'', c' c''.



Posons V - (a + b + c) = λ . On a λ = a' + a'' = b' + b'' = c' + c'' (voir la figure) c'est-à-dire que le point x_{III} est astreint à rester sur une multiplicité linéaire à 4 paramètres dans le sous-espace (a' b' c' a'' b'' c'') à 6 dimensions.

Il est donc naturel de chercher une représentation à 9 - 2 = 7 dimensions où V soit la somme des 7 coordonnées. C'est le cas pour

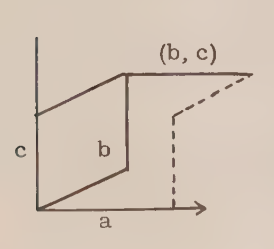
$$V = a + b + c + (ab) + (bc) + (ac) + (abc);$$

cette décomposition (parmi bien d'autres) n'est celle retenue que parce que le terme résiduel (abc) est une variance lui aussi.

4/ - Les trois familles d'axes d'information.

Il est clair qu'il y a symétrie entre tous les degrés de sondage, c'est-à-dire ici entre (a a' a'') (b b' b'') et (c c' c'').

On fera varier simplement le paramètre du 1er degré pour obtenir la 1ère famille de génératrices rectilignes; il lui correspondra une famille analogue pour chacun des autres degrés de sondage.



Faisons varier a seulement, b, c et (bc) restent constantes, pendant que a (ab) (ac) (abc) varient. Les projections des points x_{III} sur le sous-espace a, b, c, (bc), sont alignées, mais on ignore ce qui se passe dans le sous-espace (ab), (ac), (abc). Explicitons les composantes. Par exemple avec des tirages sans remise équiprobables, on a :

$$(ab) = k\left(\frac{I}{n} - \frac{I}{\nu}\right), \qquad \text{avec} \qquad a = h\left(\frac{I}{n} - \frac{I}{\nu}\right)$$

$$(ac) = k'\left(\frac{I}{n} - \frac{I}{\nu}\right),$$

$$(abc) = k''\left(\frac{I}{n} - \frac{I}{\nu}\right),$$

soit 4 fonctions linéaires de (1/n); d'où l'axe (b, c); d'où :

THEOREME -

Si la perte d'information est la variance, il existe trois familles d'axes d'information pour le sondage à 3 degrés; leur représentation métrique euclidienne demande un espace à 7 dimensions.

Corollaire.

Pour le sondage à d degrés, il y a d familles d'axes, dans l'espace euclidien à $2^d - 1$ dimensions.

Remarque. Changement de métrique.

Lorsqu'on emploie une perte d'information plus générale que la variance, la décomposition de cette perte en ses $(2^d - 1)$ composantes est invariante; et même tous les facteurs tels que $\frac{1}{\ell} - \frac{1}{\lambda}$, $\frac{1}{m} - \frac{1}{\mu}$, $\frac{1}{n} - \frac{1}{\nu}$, $\vartheta(\nu) - \vartheta(n)$ ne sont pas modifiés.

V - SUR LES PROBLEMES DE REPARTITION OPTIMUM -

1/ - Généralités.

La répartition optimum (au sens de Neyman) de l'échantillon est celle qui (pour un effectif ou un coût de sondage donné) rend minimum la variance d'échantillonnage d'un paramètre privilégié. Plus généralement on peut songer à rendre minimum une certaine perte d'information.

Tant qu'on déduit celle-ci de la variance en substituant à $(x_i - x_j)^2/2$ une fonction $f(x_i\ x_j \ldots)$ on est ramené aux problèmes classiques, moyennant quelque changement de paramètre; par exemple, au lieu d'effectifs de strate proportionnels aux

$$\sqrt{\nu_h\ C_h^2/\nu_h - 1}$$

on les prendrait proportionnels aux $\sqrt{\gamma_h}$, dans le cas du sondage stratifié (Neyman 1934).

Nous avons montré ailleurs (T.5, Ch. 3) comment à σ_h pouvaient être substitués d'autres paramètres, quand le but immédiat du sondage était d'estimer, non plus un certain ξ

$$\xi = \Sigma_i \Sigma_i x_{hi}$$

mais par exemple les σ_h eux-mêmes (en vue d'estimer ξ par un autre sondage).

2/ - Un problème nouveau.

Examinons ici le cas où l'on estime (avec biais) a la plus grande valeur de X; les cas de la plus petite valeur et de l'étendue sont analogues. La perte d'information sera définie comme ci-dessus (Chapitre IV, § II, 4) pour une strate

$$[a - E\hat{X}]_h \qquad \text{(strate h)}$$

Soit $\hat{\hat{X}}$ le plus grand des $\hat{X}$ de strate, c'est-à-dire la plus grande valeur de l'échantillon. La perte d'information correspondante est manifestement :

$$p = a - \mathcal{E}\hat{\hat{X}} = a - \Sigma \varpi_h E \hat{X}_h$$

ϖ_h étant la probabilité que le plus grand des X soit dans la strate h.

La stratification optimum serait celle qui (pour un coût $\Sigma_h n_h C_h$ ou un effectif $\Sigma_h n_h$ donné) provoquerait la perte d'information minimum. Pratiquement on connaît une variable Y en corrélation étroite avec X et on stratifie suivant des tranches de valeurs de Y. Si $\hat{\hat{X}}$ a de fortes chances de se trouver dans la strate 1, on "gonflera" l'échantillon de la strate 1; mais on hésitera à priver les autres strates de tout échantillon.

3/ - La solution.

Il est possible de montrer que l'effectif optimum n_h par strate serait, en première approximation :

a) proportionnel à la racine carrée d'un certain écart moyen entre les x_{hi};

b) inversement proportionnel à la racine carrée du coût moyen d'enquête par unité de sondage;

c) proportionnel à la racine carrée de l'effectif ν_h de la strate;

d) proportionnel à la racine carrée de la probabilité ϖ_h.

Le calcul est facile quand on suppose les n_h petits à côté des ν_h, ou les n à côté des ν (l'indice h étant sous-entendu) : la loi de distribution du plus grand $\hat{X}$ des X échantillon est $[F(X)]^n = G(\hat{X})$ où F prend

elle-même les valeurs $\left[0, \frac{1}{\nu}, \frac{2}{\nu} \ldots \frac{\nu - 1}{\nu}, 1\right]$ pour les valeurs croissantes $z_1\ z_2 \ldots z_{\nu-1},\ z_\nu = a$ de la variable X. $G(\hat{X})$ prend donc les valeurs en palier

$$0, \left(\frac{1}{\nu}\right)^n, \left(\frac{2}{\nu}\right)^n \ldots \left(\frac{\nu - 1}{\nu}\right)^n, 1$$

D'où

$$a - E\hat{X} = (z_\nu - z_{\nu-1})\left(1 - \frac{1}{\nu}\right)^n + (z_{\nu-1} - z_{\nu-2})\left(1 - \frac{2}{\nu}\right)^n + \ldots + (z_2 - z_1)\left(1 - \frac{\nu - 1}{\nu}\right)^n$$

Remplaçons $\left(1 - \frac{j}{\nu}\right)^n$ par $e^{\frac{-nj}{\nu}} = u^j$ (en première approximation).

Admettons les $(z_j - z_{j-1})$ tous de l'ordre de δ ; posons $f = n/\nu$, $u = e^{-f}$; il vient

$$a - E\hat{X} \sim \delta . (u + u^2 + \ldots + u^{n-1}) \sim \delta . u(1 - u)^{-1}$$

Le minimum de la perte d'information (toutes strates)

$$\Sigma_h \left[\varpi \frac{\delta\ u}{1 - u}\right]_h$$

lié par $\Sigma_h [c\ \nu\ f]_h = c_o$

est donné par $(1 - u)^2 \div u \left.\frac{\varpi\ \delta}{c\ \nu}\right]_h$

d'où approximativement $f^2 \div \left.\frac{\varpi\ \delta}{c\ \nu}\right]_h$

4/ - Evaluation des ϖ_h.

En réalité les ϖ_h sont très différents les uns des autres lorsque la stratification est faite efficacement. Il serait peu réaliste de n'en pas tenir compte sous prétexte qu'on ne connaît pas leurs valeurs.

Lorsqu'on se sent capable de donner des valeurs numériques subjectives $q_1 \ldots q_h$, aux probabilités a priori que $a_1 \ldots a_h$ soit le plus grand des a, il résulte de la formule de Bayes que ϖ_h et q_h diffèrent peu l'un de l'autre, si la stratification est bien faite; l'emploi de coefficients arrondis 0,8; 0,5; ou 0,1 est très suffisant en pratique.

CHAPITRE VI

SONDAGES DIVERS

I - SONDAGES A DEUX PHASES -

A. 1/ - Généralités.

On dit que le sondage est à deux phases lorsque l'enquête porte sur un caractère auxiliaire Y soumis à un sondage double (1ère et 2ème phases) - et sur les caractères principaux X soumis au sondage à la seconde phase seulement de l'enquête. Ceci constitue un procédé pratique d'enquête (sociale par exemple) pourvu qu'il existe entre X et Y une forte corrélation, et que le coût des opérations d'enquête concernant Y sans X soit très faible comparé à celui concernant à la fois X et Y. Par exemple ce sera le cas lorsque Y est obtenu en consultant des documents, alors qu'il faut envoyer un enquêteur sur le terrain pour recueillir la valeur de X.

2/ - Cas où l'échantillon de la 2ème phase est extrait du grand échantillon de la 1ere phase.

Supposons que la 2ème phase porte sur un petit échantillon prélevé dans le grand échantillon tiré à la 1ère phase. Entre les deux phases on peut stratifier le grand échantillon suivant les valeurs de Y, et le petit est extrait de cette sous-population stratifiée; cette méthode est due à Neyman (1938)[1].

Soit $\overline{X}_2$ l'estimateur de $\overline{x}$ après les deux phases, soit $\overline{X}_1$ l'estimateur qu'on pourrait former pour $\overline{x}$ avec le grand échantillon, si l'on recueillait les données x_i correspondantes (ce qui n'est justement pas le cas). Admettons que $\overline{X}_1$ et $\overline{X}_2$ soient sans biais; il vient[2] :

$$\mathcal{V}\overline{X}_2 = \mathcal{E}(\overline{X}_2 - x)^2$$

$$= \mathcal{E}(\overline{X}_2 - \overline{X}_1)^2 + \mathcal{E}(\overline{X}_1 - \overline{x})^2 + 2\ \mathcal{E}(\overline{X}_2 - \overline{X}_1')(\overline{X}_1 - \overline{x})$$

----

(1) N2. La théorie de Neyman se limite à des tirages bernoulliens.

(2) Ce calcul est emprunté à T. 4.

Mais on a :

$$\mathcal{E}\,[(\overline{X}_2 - \overline{X}_1)\,.\,(\overline{X}_1 - \overline{x})] = 0$$

le tirage du petit échantillon ne dépendant pas des résultats de la 1ère pnase.

D'autre part, pour un grand échantillon déterminé, on a :

$$E_1(\overline{X}_2 - \overline{X}_1)^2 = V_1\overline{X}_2$$

la lettre V désignant une variance aléatoire contrairement à la lettre $\mathcal{V}$. Finalement il vient :

$$\mathcal{V}\ \overline{X}_2 = \mathcal{V}\ \overline{X}_1 + \mathcal{E}(V_1\overline{X}_2)$$

a) $\mathcal{V}\ \overline{X}_1$ dépend de l'effectif du grand échantillon; par exemple si on le tire par simple tirage au sort (bernoullien ou exhaustif) $\mathcal{V}\,\overline{X}_1$ est fonction de la taille m dudit échantillon :

$$\mathcal{V}\ \overline{X}_1 = \frac{\sigma^2}{m} \quad \text{ou} \quad \frac{\sigma^2}{m}\ \frac{\mu - m}{\mu - 1}$$

μ étant l'effectif de la population.

b) $V_1\overline{X}_2$ dépend des effectifs du grand et du petit échantillons; par exemple, si le grand échantillon est découpé en strates d'effectif M_h, où les tirages effectués sont exhaustifs, on a :

$$V_1\overline{X}_2 = \sum_h \left(\frac{M_h}{m}\right)^2 \frac{M_h\,S_h^2}{M_h - 1}\left(\frac{1}{n_h} - \frac{1}{M_h}\right)$$

avec $\Sigma M_h = m$

n_h = effectif du petit échantillon, strate h.

S_h^2 = variance du grand échantillon, strate h.

On en déduit :

$$\mathcal{E}(V_1\ \overline{X}_2) = \sum_h \mathcal{E}\left[\frac{M_h^2}{m^2}\left(\frac{1}{n_h} - \frac{1}{m_h}\right)\right]\mathcal{E}\left[\frac{M_h\ S_h^2}{M_h - 1}\right]$$

en s'appuyant sur le fait que les deux crochets ont une covariance nulle, parce que, quel que soit M_h, l'espérance mathématique du dernier crochet est : $\mu_h\,\sigma_h^2/\mu_h - 1$

Si l'on désigne alors par p_h la probabilité de tirer une unité de la

strate h lors du tirage initial, on a finalement :

$$\mathcal{E}(M_h) = m\,p_h ; \quad \mathcal{E}(M_h^2) = (m\,p_h)^2 + m\,p_h q_h \frac{\mu - m}{\mu - 1}$$

$$\mathcal{E}(V_1 \ \overline{X}_2) = \sum_h \frac{\mu_h \sigma_h^2}{\mu_h - 1}\left[\frac{p_h^2}{n_h} - \frac{p_h}{m} + \frac{\mu \ p_h q_h}{(\mu - 1) n_h}\left(\frac{1}{m} - \frac{1}{\mu}\right)\right]$$

Remarque 1.

Pour $m = \mu$, on retrouve la variance du sondage stratifié limite

$$\lim \mathcal{E}\, V_1 \overline{X}_2 = \mathcal{V}\ \overline{X}_s = \sum_h \left(\frac{\mu_h}{\mu}\right)^2 \frac{\mu_h \ \sigma_h^2}{\mu_h - 1}\left(\frac{1}{n_h} - \frac{1}{\mu_h}\right)$$

Remarque 2.

Ce calcul suppose d'ailleurs n_h inférieur ou au plus égal à M_h. Or si l'on se donne à l'avance m et les n_h (avec $\sum_h n_h < m$) il y a souvent une probabilité non nulle (et parfois assez grande) que n_h dépasse M_h.

Le calcul n'est donc correct que dans des cas limités, où m reste en fait voisin de μ.

Mais il faut distinguer deux possibilités :

$M_h = 0$: Cas extrême où la strate (h) ne serait pas représentée du tout dans l'échantillon; si μ_h est très petit, l'inconvénient pratique est minime; si μ_h n'est pas petit, le cas ne peut se produire qu'exceptionnellement (si m est lui-même assez grand).

$0 < M_h < n_h$: (cas qui disparaît d'ailleurs si l'on remplace les tirages exhaustifs par des tirages bernoulliens à la deuxième phase). La contribution de la strate à la variance doit être considérée comme nulle (et non pas négative). L'effectif réel de la strate en 2ème phase est ramené de n_h à M_h (économie d'argent et perte d'information). Moyennant quoi, tout rentre dans l'ordre(1).

Dans la suite, on fera tous les calculs formellement (avec des

----

(1) Nota - Se donner m et choisir les n_h, une fois connus les M_h, ou encore se donner les n_h et augmenter m jusqu'à ce que tous les M_h soient supérieurs aux n_h, sont aussi des procédés possibles mais défectueux car : les estimateurs employés ne sont plus sans biais; la variance voit son expression modifiée.

tirages exhaustifs aux deux phases) sans chercher à savoir à partir de quel moment on sort du domaine des choses possibles.

B - VARIANCE ET PERTE D'INFORMATION.

On désire savoir quand on a le droit d'écrire :

$$\mathcal{V}\,\bar{X}'_2 = \mathcal{V}\,\bar{X}_2 + \mathcal{E}(\bar{X}'_2 - \bar{X}_2)^2$$

en envisageant deux échantillons $(m,\ n_h,\ \bar{X}_2)$ et $(m',\ n'_h,\ \bar{X}'_2)$ avec

$$m \geqslant m', \qquad n_h \geqslant n'_h$$

1/ - Etude du cas m = m'.

Alors $\mathcal{V}\,\bar{X}'_1 \equiv \mathcal{V}\,\bar{X}_1$

D'où : $\mathcal{V}\bar{X}'_2 - \mathcal{V}\bar{X}_2 \equiv E(V_1\ \bar{X}'_2) - E(V_1\ \bar{X}_2)$

c'est-à-dire : $\equiv \sum_h E\left(\frac{M_h\ S_h^2}{M_h - 1}\right) E\left(\frac{M_h}{m}\right)\left(\frac{1}{n_h} - \frac{1}{n_h}\right)$

D'autre part, $\bar{X}'_2$ étant l'estimateur $\sum_h M'_h \bar{X}'_{2h}/m$, et les (n'_h) unités étant tirées au sort parmi les (n_h), on a :

$$E_2(\bar{X}'_2 - \bar{X}_2)^2 = \sum_h \left(\frac{M'_h}{m}\right)^2 \frac{n_h\ s_h^2}{n_h - 1}\left(\frac{1}{n'_h} - \frac{1}{n_h}\right)$$

en désignant par (s_h^2) les variances de strate de l'échantillon (n_h). D'où :

$$E_1E_2(\bar{X}'_2 - \bar{X}_2)^2 = \sum_h E_1\left[\left(\frac{M'_h}{m}\right)^2 \frac{n_h\ s_h^2}{n_h - 1}\left(\frac{1}{n'_h} - \frac{1}{n_h}\right)\right]$$

$$= \sum_h \left(\frac{M'_h}{m}\right)^2 \frac{M_h\ S_h^2}{M_h - 1}\left(\frac{1}{n'_h} - \frac{1}{n_h}\right)$$

Enfin :

$$E\ E_1E_2(\bar{X}'_2 - \bar{X}_2)^2 = \sum_h E\left(\frac{M'_h}{m}\right)^2 E\left(\frac{M_h\ S_h^2}{M_h - 1}\right)\left(\frac{1}{n'_h} - \frac{1}{n_h}\right)$$

c'est-à-dire :

$$E\ E_1E_2(\bar{X}'_2 - \bar{X}_2)^2 = \mathcal{V}\bar{X}'_2 - \mathcal{V}\bar{X}_2$$

Conclusion.

Lorsqu'on maintient constant l'effectif du grand échantillon, on a le droit de considérer la variance comme une perte d'information.

2/ - <u>Représentation graphique</u>.

On peut repérer $\bar{X}_2$ en coordonnées cartésiennes (α , β) avec :

$$\alpha = \overline{\mu\, m} = \mathcal{V}(\bar{X}_1)$$

$$\beta = \overline{m\bar{X}_2} = \mathcal{E}\, V_1 \bar{X}_2$$

L'axe Oα porte les points représentatifs de la première phase du sondage ; les écarts le long de cet axe sont des pertes d'information.

Pour <u>m = Cste</u>, on vient de voir que les différences $\mathcal{V}(\bar{X}_1) - \mathcal{V}(\bar{X}_2)$, c'est-à-dire les écarts, parallèlement à l'axe des β , sont également des <u>pertes d'information</u> (en particulier pour $m = \mu$: sondage stratifié limite) : ces <u>verticales</u> (y compris l'axe des β) sont des "<u>génératrices d'information</u>".

3/ - <u>Recherche d'une seconde famille de génératrices (n_h = constantes)</u>

On va maintenir les n_h constants et faire varier m (à partir de m = μ , qui correspond à $\bar{X}_2 = \bar{X}_s$).

On a :

$$\mathcal{V}\bar{X}_1 = \frac{\mu\ \sigma^2}{\mu - 1}\left(\frac{1}{m} - \frac{1}{\mu}\right)$$

avec :

$$\sigma^2 = \sum_h p_h \sigma_h^2 + \sum_h p_h (\bar{x}_h - \bar{x})^2$$

D'où :

$$\mathcal{V}\bar{X}_2 - \mathcal{V}\bar{X}_s = \left[\frac{\mu\ \sigma^2}{\mu - 1} - \sum_h p_h \frac{\mu_h \sigma_h^2}{\mu_h - 1} + \sum_h p_h \frac{\mu_h \sigma_h^2}{\mu_h - 1}\ \frac{\mu\ q_h}{\mu - 1}\ \frac{1}{n_h}\right]\left(\frac{1}{m} - \frac{1}{\mu}\right)$$

$$= \frac{\mu}{\mu - 1}\left[\sum_h p_h(\bar{x}_h - \bar{x})^2 + \frac{\mu_h}{\mu_h - 1}\ p_h q_h \sigma_h^2\left(\frac{1}{n_h} - \frac{1}{\mu_h}\right)\right]\left(\frac{1}{m} - \frac{1}{\mu}\right)$$

4/ - <u>Conséquences</u>.

Lorsqu'on figure un sondage à deux phases par un grillage (α , β) et qu'on porte les longueurs $\alpha = \mathcal{V}\bar{X}_1$, $\beta = \mathcal{V}\bar{X}_s$ sur les deux axes, on n'a ni $\mathcal{V}\bar{X}_2 = \alpha + \beta$, ni $\alpha + \beta + \gamma$, mais :

$$\mathcal{V}\,\overline{X}_2 = \alpha + \beta - \gamma$$

avec

$$\gamma = \frac{\mu}{\mu - 1} \sum_h p_h \sigma_h^2 \left[1 - \frac{\mu_h}{\mu_h - 1}\left(\frac{1}{n_h} - \frac{1}{\mu_h}\right)\right]\left(\frac{1}{m} - \frac{1}{\mu}\right)$$

a) Il arrive que les p_h soient tous égaux, soit $p_h = p$; auquel cas on a :

$$\beta = p \sum p\ \sigma_h^2 \frac{\mu_h}{\mu_h - 1}\left(\frac{1}{n_h} - \frac{1}{\mu_h}\right)$$

d'où

$$\sigma^2\gamma = \alpha\left[\sum_h p\ \sigma_h^2 - \frac{\beta}{p}\right]$$

équation d'un paraboloïde hyperbolique dans l'espace euclidien (α, β, γ). Avec $\alpha = \beta = 0$, on retrouve l'axe des β (sondages stratifiés limites). En revanche, si $n_h = \mu_h$ quel que soit h, il vient : $\gamma = \alpha\,(\Sigma\, p\sigma_h^2 / \sigma^2)$ qui se projette sur l'axe $O\alpha$ mais n'est pas cet axe).

[D'où :

$$\mathcal{V}\,\overline{X}_2 = \alpha + \beta - \gamma = \alpha\,[\Sigma\ p(\overline{x}_h - \overline{x})^2/\sigma^2]$$

alors qu'on s'attendait à $\mathcal{V}\,\overline{X}_2 = \mathcal{V}\,\overline{X}_1 = \alpha$. En réalité, on a : $\Sigma\, n_h \equiv \mu$, donc $m = \mu$, $M_h \equiv \mu_h$, donc $\alpha \equiv 0$].

Ainsi, si $p_h = p$, il existe une image métrique euclidienne du grillage du sondage à deux phases, analogue à celle du sondage à deux degrés, mais l'information (perdue) est $\alpha + \beta - \gamma$, au lieu de $\alpha + \beta + \gamma$.

b) En général les p_h ne sont pas égaux; et l'équation en α, β, γ n'existe plus.

On retrouve le même phénomène que pour le sondage à deux degrés (où en dehors du cas où $\nu_h = \overline{\nu}$, l'image euclidienne demande des conventions adéquates).

On peut toujours poser :

$$\beta = \sum_h \beta_h, \qquad \beta' = \sum_h \beta_h/p_h, \qquad b = \sum_h p_h\sigma_h^2$$

avec :

$$\beta_h = p_h^2\ \frac{\mu_h\sigma_h^2}{\mu_h - 1}\left(\frac{1}{n_h} - \frac{1}{\mu_h}\right)$$

'où $$\sigma^2\gamma = \alpha\,(b - \beta')$$

quation d'un paraboloïde hyperbolique; les génératrices (α = cons-antes) correspondent à m constant; celles (β' = constantes) corres-ondent aux n_h constants. Mais la perte d'information symétrique

$\mathcal{V}\,\overline{X}$ est $\alpha + \beta - \gamma$ et non $\alpha + \beta' - \gamma$

Avec L (= 1/p) strates, il faut utiliser (L + 2) dimensions (α β_h γ) t une surface d'équation

$$\sigma^2\gamma = \alpha\,(b - \Sigma\, \beta_h/p_h)$$

'où la perte $$\mathcal{V}\,\overline{X}_2 = \alpha + \Sigma\, \beta_h - \gamma$$

)n voit facilement que $\mathcal{E}(\overline{X}_s - \overline{x})\,(\overline{X}_2 - \overline{X}_s) > 0$

$$\mathcal{E}\,(\overline{X}_1 - \overline{x})\,(\overline{X}_2 - \overline{X}_1) > 0$$

Peut-on parler du gain d'information dû à une stratification complète ?

Lorsqu'on peut stratifier toute la population, on ne fait pas de ondage à deux phases : au sondage stratifié d'effectifs n_h correspond a perte d'information $\mathcal{V}\,\overline{X}_s$. On vérifierait que $\mathcal{V}\,\overline{X}_2 - \mathcal{V}\,\overline{X}_s = \alpha - \gamma$ est po-itif. On peut considérer que cette différence mesure l'information ga-née à stratifier, ou plutôt la perte d'information (à partir de $\overline{X}_s$) due une stratification ne portant que sur m unités (au lieu de μ).

Ce type de question sera examiné au Chapitre VII, IIème partie.

I - INDICATIONS SUR QUELQUES AUTRES PLANS -

On a signalé au Ch. III l'intérêt qu'il pouvait y avoir parfois à fi-gurer une famille de plans de sondage sur un treillis en quelque sorte léthorique; d'autre part on connaît la forme générale des pertes d'in-ormation (sous certaines hypothèses restrictives, Ch. IV); disons un not de la variance d'échantillonnage d'estimateurs sans biais relatifs des méthodes de sondage (assez courantes) sortant plus ou moins du adre du chapitre V.

1/ - Sondages stratifiés "optimum".

Soit une suite de plans de sondage stratifiés dont chacun est "op-imum" au sens de Bowley, Neyman ou Yates, l'effectif ou le coût dé-croissant régulièrement; on peut représenter cette suite en portant $\mathcal{V}\,\overline{X}$ sur un axe (voir Ch. V, fin du § II).

On peut également repérer les points (opt n_h) sur le treillis des sondages stratifiés correspondants, sommets d'une "trajectoire" en ligne brisée qui n'est pas un axe d'information (voir Ch. III, § II. 5.b).

$\mathcal{V}\,\overline{X}$ est combinaison linéaire des pertes par strate et vérifie la relation de récurrence.

2/ - Stratification a posteriori.

Sur le treillis précédent on envisage des "surfaces d'onde" ($\Sigma\, n_h = n$); sur chacune d'elles on donne une distribution de probabilité pour l'estimateur $\overline{X}_s$ et la variance $V\overline{X}_s$; la taille n de l'échantillon étant donnée, les n_h sont aléatoires (tirage au sort de l'échantillon dans une urne où toutes les strates sont mélangées). On pose $\mathcal{E}\, V\overline{X}_s = p(\overline{X})$.

Si certains des n_h sont nuls ($\overline{X}_h$ indéterminés), on convient à l'avance de l'expression à donner à $\overline{X}_s$ pour qu'il ne soit pas indéterminé (sans introduire de biais).

Ici encore il est clair que $p(\overline{X})$ est combinaison linéaire de pertes par strate $\mathcal{E}(V\overline{X}_h)$ et vérifie la relation de récurrence.

CHAPITRE VII

LES INFORMATIONS SUPPLÉMENTAIRES

I - LE CHANGEMENT D'ESTIMATEUR -

A - GENERALITES.

1/ - Rappel de la technique classique.

On sait que, pour un échantillon d'ores et déjà tiré, le choix d'un estimateur plus efficace que l'estimateur courant (construit avec le seul plan d'échantillonnage : probabilités de tirage, effectifs, etc.) est rendu parfois possible grâce à des "informations supplémentaires". Les cas habituels sont :

- l'estimation par ratio (ou par quotient);
- l'estimation par une formule de régression;
- l'estimation par une formule de stratification (a postériori).

Ces dernières années Žarković a ajouté à cette liste l'estimation par différence (Réf. Z1).

Dans tous ces cas, on se propose d'estimer pour une population donnée une certaine moyenne $\overline{x}$ d'une variable X, alors qu'on a des informations sur une variable Y en corrélation étroite avec X. Rien n'empêche de substituer à $\overline{x}$ et X des expressions ζ et Z plus générales.

Même lorsqu'il est question de strates, elles n'interviennent qu'après coup. Aussi peut-on supposer (pour simplifier) que le sondage est à un seul degré, sans strate, avec probabilités égales (treillis réduit à un axe, commun à tous estimateurs).

On connaît, non seulement les x_i des unités-échantillon, mais aussi leurs y_i, ainsi que la moyenne $\overline{y}$ pour la population entière, et plus généralement toute la distribution des y_i.

On indiquera d'autres cas où des renseignements d'une nature différente permettent de réduire la variance d'échantillonnage en changeant d'estimateur.

On peut rapprocher enfin cette question de celle de l'estimation du maximum de vraisemblance et plus généralement du recours à des hypothèses sur la loi mathématique qui règlerait la distribution des x_i, hypothèses permettant de changer d'estimateur et (par là) de réduire la variance.

2/ - Problème.

A-t-on le droit d'appeler gain d'information la réduction de variance ? Autrement dit : s'agit-il bien d'une quantité d'information ?

Si Z et Z' sont deux estimateurs sans biais de ζ, avec $\mathcal{V}Z > \mathcal{V}Z'$, il est exact que $\mathcal{V}Z - \mathcal{V}Z'$ est une perte d'information générale (au sens des ch. II et III). Mais il n'est pas en général exact que ce soit égal à $\mathcal{E}(Z - Z')^2$.

Or, il importe assez peu que $(\mathcal{V}Z - \mathcal{V}Z')$ soit une information dans l'optique d'une suite de plans de sondage d'une famille F, alors qu'on a changé d'optique et qu'on n'envisage qu'un seul plan (et même un seul échantillon). Ce n'est plus en augmentant la taille de l'échantillon qu'on passe du point Z au point Z'. Il n'en serait pas moins intéressant de pouvoir représenter ZZ' et ζ par trois points alignés, $\mathcal{V}Z - \mathcal{V}Z'$ étant la distance (ZZ'), mesurée avec la même métrique que (ζ Z) ou (ZZ'), c'est-à-dire $\mathcal{E}(Z - Z')^2$.

On dirait alors que $(\mathcal{V}Z - \mathcal{V}Z')$ est l'information supplémentaire.

Dans les divers cas étudiés s'est retrouvé le résultat que voici.

THEOREME -

Si l'estimateur Z' est substitué à Z, l'un et l'autre sans biais, avec $\mathcal{V}Z' < \mathcal{V}Z$, $\mathcal{V}Z - \mathcal{V}Z'$ n'est pas en général une information supplémentaire, à moins qu'on n'ait :

$$\mathcal{E}(Z' - \zeta)(Z - Z') = 0 ;$$

En effet, si l'on convient de poser, comme définition de l'information supplémentaire :

$$\mathcal{V}Z - \mathcal{V}Z' = \mathcal{E}(Z - Z')^2 ,$$

c'est cette relation qui exprime l'orthogonalité des deux vecteurs aléatoires $\overline{\zeta Z'}$ et $\overline{Z'Z}$.

Suivant les problèmes traités, Z' est bien déterminé ou au con-
raire dépend de paramètres, par exemple linéairement (auquel cas on onsidèrera un point Z' décrivant une multiplicité linéaire). Nous re-viendrons plus loin sur le cas où Z' est biaisé.

3/ - Exemple.

Supposons tiré un échantillon d'effectifs n_h dans chaque strate notations du Ch. V). Parmi tous les estimateurs linéaires sans biais $Z_\lambda = \Sigma\, \lambda_h \overline{X}_h$ dépendant des λ_h liés par $\Sigma_h \lambda_h \overline{x}_h = \overline{x}$, l'estimateur de variance minimum Z° serait tel que $d\,[\mathcal{V} Z_\lambda + \mu \Sigma\, \lambda_h \overline{x}_h] = 0$ (μ de Lagrange) e qui signifie que les λ_h devraient être proportionnels aux $n_h \overline{x}_h / \tau_h^2$, avec $\tau_h^2 = \nu_h \sigma_h^2 / \nu_h - 1$.

Les informations supplémentaires à connaître seraient ici les $\tau_h^2 / \overline{x}_h$ ou (si tirages bernoulliens) les $\sigma_h^2 / \overline{x}_h$.

Pratiquement (depuis Bowley) on pose $\lambda_h = \nu_h / \nu$ pour éliminer e biais avec des poids indépendants de la variable X étudiée, en l'ab-sence d'informations supplémentaires; posons $Z = \Sigma\, \nu_h\, \overline{X}_h / \nu$;

$$\mathcal{V} Z - \mathcal{V} Z^\circ \qquad\qquad (\geqslant 0)$$

représente l'information perdue faute de connaître les $\tau_h^2 / \overline{x}_h$; alors que

$$\mathcal{V} Z_\lambda - \mathcal{V} Z \qquad\qquad (\gtrless 0)$$

ne devrait pas être tenu comme un gain ou une perte d'information faute d'user des poids (ν_h / ν).

Remarque.

Supposons les n_h choisis (à la Neyman) proportionnellement aux $\tau_h \nu_h / \sqrt{C_h}$) d'une certaine variable X. L'estimateur de Neyman cor-respond à $\lambda_h = \nu_h / \nu$; alors que l'estimateur de variance minimum cor-respond à λ_h proportionnel à

$$\nu_h \overline{x}_h / \tau_h \sqrt{C_h}$$

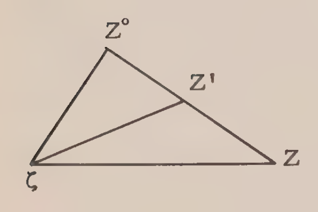
a coïncidence exigeant que le coût moyen d'enquête par strate C_h soit nversement proportionnel à $(\nu_h \gamma_h^2 / \nu_h - 1)$, en gros à γ_h^2, carré du oefficient de variation par strate.

Ceci est à noter, d'autant plus que l'optimum au sens de Neyman e concerne jamais qu'une variable particulière X; or cet "optimum" ninimise la variance d'un estimateur très particulier:celui qui reste ans biais pour toutes variables X; après quoi un estimateur Z° devrait tre substitué à Z.

4/ - Retour au cas général.

1ère interprétation.

Dans l'espace euclidien auxiliaire où $\overline{\zeta Z}^2$ représente $\mathcal{V} Z$, figurons par Z' l'estimateur de ζ obtenu grâce aux informations supplémentaires.



Supposons $\mathcal{V} Z' < \mathcal{V} Z$

mais $\mathcal{V} Z - \mathcal{V} Z' \neq \mathcal{E}(Z' - Z)^2$

Soit Z^0 la projection de ζ sur ZZ', c'est-à-dire $\mathcal{V} Z^0$ le minimum de $\mathcal{V} Z'$.

$$\mathcal{V} Z' = \mathcal{E}(Z^0 - \zeta)^2 + \mathcal{E}(Z' - Z^0)^2$$

$$\mathcal{V} Z = \mathcal{E}(Z^0 - \zeta)^2 + \mathcal{E}(Z - Z^0)^2$$

Par analogie avec la relation de récurrence de l'information d'après Schutzenberger, on dira que :

$$\mathcal{E}(Z' - Z^0)^2 \text{ ou } \mathcal{E}(Z - Z^0)^2$$

est une perte d'information par rapport à l'estimateur Z^0.

En revanche,

$$\mathcal{V} Z - \mathcal{V} Z' = \mathcal{E}(Z - Z^0)^2 - \mathcal{E}(Z' - Z^0)^2$$

ne sera pas considéré comme une information supplémentaire.

2ème interprétation.

Dans l'espace euclidien auxiliaire où ζZ représente $\mathcal{V} Z$, les points ζ Z^0 Z' Z sont alignés. La distance $Z'Z$ est mesurée par $\mathcal{V} Z - \mathcal{V} Z'$ mais non par $\mathcal{E}(Z - Z')^2$, à moins que Z' ne coïncide avec Z^0.

ζ Z⁰ Z' Z

On ne considèrera donc pas la droite ζ Z^0 Z' Z comme un véritable axe d'information(1).

----

(1) Construire un triangle $\zeta Z Z'$ avec $\zeta Z = \mathcal{V} Z$, $\zeta Z' = \mathcal{V} Z'$, $ZZ' = \mathcal{E}(Z - Z')^2$ n'est possible que si le coefficient de corrélation entre Z et Z' est compris entre 0 et une limite supérieure au plus égale à 1 (cas $\mathcal{V} Z = \mathcal{V} Z'$).(espace euclidien).

B - ETUDE DE CAS OU LES DEUX ESTIMATEURS SONT SANS BIAIS.

On retrouve le triangle rectangle $\zeta Z^0 Z$ ou plutôt $\overline{x}X^0\overline{X}$ dans les cas suivants :

1/ - Estimation par différence.

L'estimateur par différence (de Žarković) X est tel que :

$$X^0 - \overline{X} = \overline{y} - \overline{Y}$$

Il n'y a ici orthogonalité que dans un cas particulier :

$$\rho_{xy} = \rho = \sigma_y/\sigma_x$$

bien que la variance soit réduite dès que :

$$\rho > \sigma_y / 2\sigma_x$$

Il suffit d'écrire la relation bien connue :

$$\rho\, \sigma_x = b\,\sigma_y$$

où b est la pente de la droite de régression de x en y, pour voir que la structure particulière pour laquelle il y a gain d'information au sens strict est celle où

$$\underline{b = 1}$$

2/ - Estimation par ratio : $X^0 = \overline{X}\,\overline{y}/\overline{x}$

On supposera d'emblée que le biais est négligeable (si non nul), c'est-à-dire :

$$\rho\, \gamma_x - \gamma_y = 0$$

c'est-à-dire $b = \overline{x} / \overline{y}$

qui signifie que la droite de régression de X en y passe par l'origine. C'est le cas où la variance $\mathcal{V} X^0$ diffère très peu de :

$$\mathcal{V}\overline{X}\left[1 - \rho^2 + \left(\frac{\gamma_y}{\gamma_x} - \rho\right)^2\right] \equiv \mathcal{V}\,\overline{X}(1 - \rho^2)$$

(égalité rigoureuse si la distribution XY a la structure du N°3).

3/ - Estimation par régression : $X^0 = \overline{X} + b(\overline{y} - \overline{Y})$.

On s'en tient au cas "idéal" où b lui-même est connu (et non une estimation B sur échantillon) tout en postulant que la distribution à deux dimensions $(x_i\ y_i)$ a la structure particulière suivante :

- la distribution des y_i est quelconque;
- la droite de régression de X en y est quelconque : $x = a + by$;
- on a $X = a + by_i + \sigma_i\vartheta$ avec $\mathcal{E}\,\vartheta = 0$; $\mathcal{E}\,\vartheta^2 = 1$;
- la distribution de ϑ est indépendante de celle de Y;
- σ_i dépend de y_i.

On a alors à nouveau $\mathcal{V}X^0 = \mathcal{V}\overline{X}\,(1 - \rho^2)$.

4/ Sondage bernoullien avec identification.

On désignera ainsi le cas où les boules portent un signe distinctif (un numéro d'identité) permettant de reconnaître, parmi n boules tirées, celles qui ont été tirées 2 fois, 3 fois, ...

Alors on a intérêt à choisir comme estimateur, au lieu de la moyenne X_n, le rapport suivant :

$$X_n^0 = \frac{\text{total des x tirés (chacun compté une fois)}}{\text{nombre de boules distinctes tirées}}$$

Cet estimateur, qui s'identifie (avec des probabilités diverses) à ceux des sondages exhaustifs d'effectif n, n - 1, n - 2, ... est évidemment sans biais; et sa variance est une combinaison linéaire des variances desdits estimateurs exhaustifs, soit :

$$V_n, \qquad V_{n-1}, \qquad V_{n-2}, \ldots$$

On trouve ainsi :

$$\mathcal{V}X_2^0 = \frac{1}{\nu}V_1 + \frac{\nu - 1}{\nu}V_2$$

$$\mathcal{V}X_3^0 = \frac{1}{\nu^2}V_1 + 3\,\frac{\nu - 1}{\nu^2}V_2 + \frac{(\nu - 1)(\nu - 2)}{\nu^2}V_3$$

$$\mathcal{V}X_4^0 = \frac{1}{\nu^3}V_1 + 7\,\frac{\nu - 1}{\nu^3}V_2 + 6\,\frac{(\nu - 1)(\nu - 2)}{\nu^3}V_3 + \frac{(\nu - 1)(\nu - 2)(\nu - 3)}{\nu^3}V_4$$

$$\mathcal{V}X_5^0 = \frac{1}{\nu^4}V_1 + 15\,\frac{\nu - 1}{\nu^4}V_2 + 25\,\frac{(\nu - 1)(\nu - 2)}{\nu^4}V_3 + 10\,\frac{(\nu - 1)\ldots}{\nu^4}V_4 + \frac{(\nu - 1)\ldots}{\nu^4}V_5$$

Le coefficient de V_i sur la ligne $\mathcal{V} X_n^0$ est la probabilité de tirer i boules distinctes en n tirages. D'où (en explicitant) :

$$\mathcal{V} X_2^0 = \frac{\sigma^2}{2} = \mathcal{V} \bar{X}_2$$

$$\mathcal{V} X_3^0 = \frac{\sigma^2}{3} \frac{2\nu - 1}{2\nu} = \frac{\sigma^2}{3} - \frac{\sigma^2}{6\nu} < \mathcal{V} \bar{X}_3$$

$$\mathcal{V} X_4^0 = \frac{\sigma^2}{4} \frac{\nu(\nu - 1)}{\nu^2} = \frac{\sigma^2}{4} - \frac{\sigma^2}{4\nu} < \mathcal{V} \bar{X}_4$$

$$\mathcal{V} X_5^0 = \frac{\sigma^2}{5} - \frac{\sigma^2}{5} \frac{9\nu^2 - \nu - 1}{6\nu^3} < \mathcal{V} \bar{X}_5$$

et ainsi de suite indéfiniment. La formule générale est :

$$\mathcal{V} X_n^0 = \frac{1}{\nu^{n-1}} [V_1 + a_2(\nu - 1)V_2 + a_3(\nu - 1)(\nu - 2)V_3 + \ldots + (\nu - 1)\ldots(\nu - n + 1)V_n]$$

avec : $a_2 = 2^{n-1} - 1$

$a_3 = (3^{n-1} - 2^n + 1)/2!$

$a_4 = (4^{n-1} - 3^n + 3.\ 2^{n-1} - 1)/3!$ etc.

Le fait que les boules portent un signe d'identité apporte bien une information, c'est-à-dire que la différence des deux variances

$$\mathcal{V} \bar{X}_n - \mathcal{V} X_n^0$$

est égale à $\mathcal{E}(\bar{X}_n - X_n^0)^2$, autrement dit : $\bar{x} X_n^0$ et $\bar{X}_n X_n^0$ sont orthogonaux. En effet :

Raisonnons dans le cas où n = 4. Supposons qu'on ait tiré (a b c d), on a $\bar{X} = X^0$. Si on a tiré au contraire (a c c d), on a $X^0 = \frac{a + c + d}{3}$ Mais à cet X^0, correspondent trois $\bar{X}$ distincts, équiprobables :

$$\frac{2a + c + d}{4}, \qquad \frac{a + 2c + d}{4}, \qquad \frac{a + c + 2d}{4}$$

On a évidemment :

$$\mathcal{E}(\bar{X} \mid X^0) = \frac{1}{3} \frac{4a + 4c + 4d}{4} = X^0$$

d'où suit l'orthogonalité, et le gain d'information.

X^0 est la projection de $\overline{x}$ sur la multiplicité linéaire

$$\lambda x_i + \mu x_j + (1 - \lambda - \mu)\overline{X}$$

en désignant par x_i, x_j, les x des boules tirées plusieurs fois.

5/ - Estimation du carré de l'écart-type (avec ou sans la vraie moyenne).

La question suivante nous a été posée (1956) par M. Fonsagrive, pour les sondages bernoulliens.

On possède un échantillon, avec lequel on estime le σ^2 de l'urne-mère. Quelle information m'apporte-t-on si l'on m'apprend par ailleurs la vraie valeur de la moyenne de la population ? (plus correctement : de combien est réduite la perte d'information ?)

a) Si $\overline{x} = \mu_1$ est connu, on a un échantillon de n valeurs

$$z_i = (x_i - \mu_1)^2$$

dont la moyenne $\overline{Z} = S_i(x_i - \mu_1)^2/n$

estime sans biais $\zeta = \sigma^2$, avec la variance (où σ_z^2 désigne $\mu_4^0 - \sigma^4$)

$$\mathcal{V}\overline{Z} = \begin{cases} \sigma_z^2/n & \text{tirages avec remise} \\ \dfrac{\sigma_z^2}{n}\,\dfrac{\nu - n}{\nu - 1} & \text{tirages sans remise} \end{cases}$$

b) Si $\overline{x} = \mu_1$ n'est pas connu, mais estimé par $\overline{X}$, on estime σ^2 par

$$\frac{\nu - 1}{\nu}\,\frac{n\,S^2}{n - 1} = \frac{\nu - 1}{\nu}\overline{Y}$$

1) Faisons tendre ν vers l'infini; on sait, cf. Ch. I, § III 8 a (cas du sondage bernoullien) que :

$$\mathcal{V}\overline{Y} = \mathcal{V}\overline{Z} + 2\sigma^4/n(n - 1)$$

Donc $2\sigma^4/n(n - 1)$ est la réponse à la question posée; sous réserve qu'il soit bien justifié d'appeler "information" cette réduction de variance.

2) Si ν est fini, on sait (cf. Ch. I, § III 8 b) que $\mathcal{V}[(\nu - 1)\overline{Y}/\nu]$ est de la forme :

$$A.\mathcal{V}\overline{Z} + B\sigma^4$$

et on montre facilement que $A > 1$, $B > 0$ (dès que $\nu > 2$, $n \geqslant 2$), de sorte que la connaissance de μ_1 réduit la variance de :

$$(A - 1)\mathcal{V}\,\overline{Z} + B\sigma^4$$

sans qu'on puisse considérer comme pleinement justifié l'emploi du mot "information" pour désigner cette expression.

1) Dans le cas des tirages avec remise (ou si ν est infini) on va vérifier qu'on a bien :

$$\mathcal{V}\,\overline{Y} - \mathcal{V}\,\overline{Z} \equiv \mathcal{E}\,(\overline{Y} - \overline{Z})^2$$

En effet il suffit d'établir que

$$\mathcal{V}\,\overline{Z} \equiv \mathrm{Cov}\,\overline{Y}\,\overline{Z}$$

ou encore $$\mathcal{E}\,\overline{Z}^2 \equiv \mathcal{E}\,\overline{Y}\,\overline{Z}$$

On a : $$\mathcal{E}\,\overline{Z}^2 = \mathcal{E}\left(\frac{S\,z_i^2}{n}\right)^2$$

$$\overline{Y} = \frac{n}{n-1}\;\frac{S(z_i - z_j)^2}{n^2}$$

$$\equiv \frac{S\,z_i^2}{n} - \frac{2\,S\,z_i z_j}{n(n-1)}$$

car le développement de $S(z_i - z_j)^2$ comprend $n(n-1)$ fois $(z_i\,z_j)$; donc dans $\overline{Y}$ figure une fois seulement le terme $(z_i\,z_j)$. Il vient :

$$\mathcal{E}\,\overline{Y}\,\overline{Z} = \mathcal{E}\left(\frac{S\,z_i^2}{n}\right)^2 - 2\,\mathcal{E}\left(\frac{S\,z_i^2\,.\,S\,z_i\,z_j}{n^2(n-1)}\right);$$

et l'espérance du second terme est nulle puisqu'il est de la forme $\lambda\,\mathcal{E}(z_i^3\,z_j)$, avec :

$$\mathcal{E}\,z_i^3\,z_j = \mathcal{E}\,z_i^3\;\mathcal{E}\,z_j, \qquad \mathcal{E}\,z_j = 0 \qquad \text{c. q. f. d.}$$

2) Mais dans le cas des tirages sans remise, on montrera qu'il n'en est plus ainsi, sur un simple exemple :

$$n = 2, \qquad \nu = 3, \qquad z_i = a,\ b,\ c.$$

Il vient :

$$\mathcal{E}\,\overline{Z}^2 = \left(\frac{a^2 + b^2}{2}\right)^2 + \left(\frac{b^2 + c^2}{2}\right)^2 + \left(\frac{c^2 + a^2}{2}\right)^2$$

D'autre part, en substituant $\nu \overline{Y} / \nu - 1$ à $\overline{Y}$ pour n'avoir pas de biais, on a :

$$\frac{\nu}{\nu - 1} \mathscr{E} \; \overline{Y} \; \overline{Z} = \frac{3}{2}\left[\frac{a^2 + b^2}{2} \cdot \frac{(a - b)^2}{2} + \ldots\right]$$

$$= \left(\frac{a^2 + b^2}{2}\right)^2 + \ldots + \frac{1}{8}(a^4 - 6\,a^3b + 2\,a^2b^2 - 6\,ab^3 + b^4) + \ldots$$

Et il n'y a aucune raison pour que les seconds termes, soit

$$[(a - b)^4 - 2(a + b)^2 ab]$$

s'annulent.

Conclusion.

La connaissance du premier moment apporte un gain d'information si les tirages sont bernoulliens, une réduction de variance s'ils sont exhaustifs. Le vocable "information" apparaît finalement comme assez dangereux par les interprétations abusives qu'on serait tenté de lui donner dans les comparaisons entre estimateurs. Il n'est pas toujours facile de discerner une famille d'estimateurs Z' dont Z fait partie et dont Z° est le point le plus proche de ζ ; elle nous échappe dans les cas n°4 et 5 ci-dessus.

C - COMPARAISON ENTRE ESTIMATEURS EVENTUELLEMENT BIAISES.

1/ - Généralités et Rappels.

Tant que le biais était nul (§ A et B), variance et perte d'information étaient pris l'un pour l'autre; alors qu'en fait tout multiple de la variance est perte d'information symétrique, sans oublier l'existence des pertes d'information asymétriques; ces points prennent quelque importance pour les estimateurs biaisés.

On a vu (Ch. IV, § II) que la notion de perte d'information ne s'étendait qu'aux estimateurs qu'on peut appeler uniformément biaisés; et on sait calculer les pertes d'informations correspondantes dans des cas suffisamment étendus en pratique, notamment le cas dit isomorphe.

On a rejeté par conséquent, comme insuffisamment fondée, la technique qui consiste (pour estimer ζ) à comparer la variance $\mathcal{V} Z'$ de l'estimateur Z' sans biais et la somme $(\mathcal{V} Z + b^2)$ de l'estimateur Z affecté du biais b. Peut-être cette technique avait-elle son origine dans la théorie classique de l'estimation où (dans le cas régulier) on démontre que

$$\mathcal{V} Z + b^2 = \mathcal{E}(Z - \zeta)^2$$

est supérieur (au plus égal) à

$$\left(1 + \frac{d\,b}{d\,\zeta}\right)^2 / n\,H$$

théorème que, pour $b \equiv 0$, se réduit à l'inégalité bien connue

$$\mathcal{V} Z \geqslant 1/n\,H.$$

Parmi les pertes p(Z) (où ni $\mathcal{V}Z$, ni $\mathcal{V}Z + b^2$ ne figurent), on a besoin à présent de savoir en choisir une qu'on puisse le plus valablement comparer à $\mathcal{V}Z'$; soit $\overline{p}$.

Il n'est pas évident d'abord que p doive être symétrique. La technique des sondages nous conseille même, pour l'estimation d'un ratio, de faire jouer un rôle privilégié à une expression asymétrique (Ch. IV, § II. 8).

On peut tenir pour sensé le principe que, si Z se rapproche infiniment de ζ (soit $n \longrightarrow \nu$, soit $n \longrightarrow \infty$) $\overline{p}(Z)$ devrait être celle des p(Z) qui a le contact le plus élevé avec $\mathcal{V} Z$ (principe mis en œuvre pour le ratio et la corrélation). Nous traiterons seulement deux exemples.

2/ - L'estimation par le ratio.

Par rapport aux notations du § B. 2, on a :

$$\zeta = \overline{x}, \qquad Z' = \overline{X}, \qquad Z = \overline{X}\,\overline{y}/\overline{Y}$$

$$\overline{p}(Z) = \overline{p}\left(\overline{y}\frac{\overline{X}}{\overline{Y}}\right)$$

Il est évident (mais utile à rappeler) que

$$\overline{p}(k\,Z) = k^2\overline{p}(Z).$$

Il vient donc

$$\overline{p}(Z) = \overline{y}^2\overline{p}\left(\frac{\overline{X}}{\overline{Y}}\right)$$

$$= \overline{y}^2\left(\frac{\overline{x}}{\overline{y}}\right)^2 \mathcal{E}\left(\frac{\overline{X}}{\overline{x}} - \frac{\overline{Y}}{\overline{y}}\right)^2$$

(cf. Ch. IV, § II.8).

ou en posant $\sigma^2 = \mathcal{V}\,\overline{X}$, $\sigma'^2 = \mathcal{V}\,\overline{Y}$

$\rho\,\sigma\,\sigma' = \text{Cov.}\,\overline{X}\,\overline{Y}$, $\gamma = \sigma/\overline{x}$, $\gamma' = \sigma'/\overline{y}$:

$$\overline{p}(Z) = \mathcal{V}\,Z' - \overline{x}^2\, 2\gamma\gamma'\left[\rho - \frac{\gamma'}{2\gamma}\right]$$

D'où : $\overline{p}(Z) \leqslant \mathcal{V}Z'$ si $\rho \geqslant \gamma'/2\gamma$

On retrouve en toute rigueur le résultat approximatif classique avec les variances.

De même, on a :

$$\overline{p}(Z) = \mathcal{V}\,Z'(1-\rho^2) + \sigma'^2\left(\frac{\rho\sigma}{\sigma'} - \frac{\overline{x}}{\overline{y}}\right)^2 \leqslant \mathcal{V}\,Z'(1-\rho^2)$$

Le minimum de la perte d'information est atteint quand le rayon vecteur du centre de gravité coïncide avec la droite de régression de X en y ($\rho\sigma/\sigma' = \overline{x}/\overline{y}$).

3/ - L'estimation par régression.

Si les axes sont translatés sans modifier la pente de la droite d'estimation (passant par (X, Y), il est clair que la perte $\overline{p}(Z)$ ne doit pas se modifier (idée que nous devons à M. Fonsagrive). Donc si la droite d'estimation a pour pente

$$(\overline{x}/\overline{y} = \rho\sigma/\sigma'),$$

la perte demeure : $\overline{p}(Z) = \mathcal{V}\,Z'(1-\rho^2)$

Et si cette pente est $\overline{x}/\overline{y} = t$, la perte demeure

$$\overline{p}(Z) = \mathcal{V}\,Z'(1-\rho^2) + \sigma'^2\left(t - \rho\,\frac{\sigma}{\sigma'}\right)^2$$

Supposons cette pente t elle-même aléatoire (soit T), déterminée par l'échantillon tiré; alors $\overline{p}(Z)$ est aléatoire. C'est notamment le cas si t est l'estimation(biaisée)de b par les moindres carrés

$$T = \frac{R\,S}{S'} = \frac{R\,S\,S'}{S'^2}$$

Alors $\overline{p}(Z)$ n'est plus (à proprement parler) perte d'information, non plus que son espérance.- Mais l'espérance relative à l'estimation isomorphe de t par T est calculable (Ch. IV, II, § 8) :

$$\log t = \log(\rho\sigma\sigma') - \log \sigma'^2$$

$$\frac{\Delta t}{t} = \frac{\Delta(\rho\sigma\sigma')}{\rho\sigma\sigma'} - \frac{\Delta\sigma'^2}{\sigma'^2}$$

$$\overline{p}(T) = t^2 \mathcal{E}\left[\frac{R\,S\,S'}{\rho\sigma\sigma'} - \frac{S'^2}{\sigma'^2}\right]^2$$

Alors on posera :

$$\boxed{\overline{p}(Z) = \mathcal{V}\, Z'(1 - \rho^2) + \sigma'^2\overline{p}(T)}$$

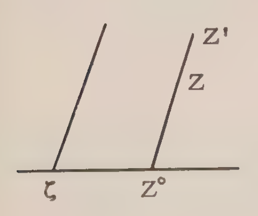
Les deux cas traités (Ratio, Régression) semblent suffire à donner une idée des simplifications que la méthode de la perte d'information introduit par rapport aux calculs classiques de variances (sans nécessiter ni approximations, ni hypothèses spéciales sur la structure des distributions).

4/ - Gain d'information est-il correct ?

On retrouve le problème discuté déjà en A et B sous une forme généralisée : p(Z) et p(Z') étant finalement des distances (ζ Z) et (ζ Z'), peut-on interpréter $|p(Z) - p(Z')|$ comme une distance (Z Z') ? Si les treillis de Z et Z' sont réduits chacun à un axe, peut-on les représenter simultanément par deux axes de même origine ζ superposés ou sécants ?

Reprenons les cas (B. 2) de l'estimation par ratio et (B. 3) de la régression; le paramètre t permet de décrire la multiplicité linéaire dont font partie $\overline{X}$ et $\overline{X}\ \overline{y}/\overline{Y}$ et où $\overline{x}$ se projette en X°; $1 - \rho^2 = \sin^2\vartheta$; ϑ est l'angle de $\overline{x}\ \overline{X}$ et $\overline{x}\ X^\circ$ (première interprétation : les pertes sont les carrés des distances euclidiennes).

Ainsi, quand les pertes sont des distances euclidiennes (2ème interprétation) on ne doit pas superposer les trois axes (Z) (Z') (Z°) de crainte d'interprétations abusives. D'ailleurs il n'y a plus de raison de vouloir considérer la distance (Z Z') comme égale à $\mathcal{E}(Z - Z')^2$ puisque (ζ Z) n'est plus $\mathcal{V}$ Z; d'ailleurs on ignore quelle expression convient pour (Z Z')



Conclusion.

Les informations supplémentaires données conduisent au partage de la perte d'information p(Z) en deux composantes dont une irréductible : p(Z°). On peut les figurer sur deux axes sécants d'origine ζ substitués à l'axe unique. On peut considérer la figure

formée d'un axe unique ζZ° et d'axes transversaux $Z^\circ ZZ'$ dont chacun correspond à un effectif donné de l'échantillon (il n'existe pas de seconde famille d'axes).

(ZZ')est un gain ou une perte d'information suivant les cas, mais ne signifie en général rien d'autre que p(Z') - p(Z).

D - CAS OU L'ON A DES INFORMATIONS SUR LA LOI THEORIQUE DE DISTRIBUTION.

1/ - Généralités.

Jusqu'ici la théorie ne comportait aucun recours à des hypothèses sur une loi théorique (par exemple laplacienne) que les variables seraient censé suivre.

Cette ignorance de la loi théorique de distribution est propre à la théorie des sondages; elle se justifie du fait que les populations sondées (au cours des enquêtes sociales, économiques, culturelles, etc.) fournissent peu d'exemples de distributions statistiques remarquables.

Beaucoup de sondages sont occasionnels et on ignore (avant d'avoir fait l'enquête) qu'on va sonder par exemple une population distribuée suivant une loi de Galton. Les enquêtes périodiques, même, offrent peu de possibilités, pour la raison qu'un questionnaire permet de recueillir sur chaque unité (i) des données X_i Y_i Z_i T_i etc. dont le nombre dépasse parfois la centaine : si le montant X des revenus de l'unité statistique (le ménage) suit une loi de Galton, il y a peu de chances que Y, Z, T, etc. soient dans le même cas.

On évite donc de faire un plan de sondage conçu en vue d'une variable X privilégiée. Mais il est tout indiqué d'employer les informations qu'on peut avoir sur la structure de la variable X, à construire des estimateurs spéciaux pour les paramètres de la loi de distribution de X. On ne le fait guère : et d'abord parce que l'emploi d'un plan de sondage stratifié, à plusieurs degrés, etc. bien adapté rend les calculs difficiles et d'ailleurs peu rentables.

Toutefois Aitchison l'a fait (Réf. UTTING-COLE, 4ème partie) pour une distribution présumée être de Galton, en appliquant un résultat de Finney (Réf.) à un sondage de l'Institut de Statistique d'Oxford (sondage supposé bernoullien).

2/ - Cas où l'échantillon a un grand effectif.

En admettant le plan de sondage simplifié ainsi à l'extrême, il convient encore de distinguer le cas où l'échantillon est assez grand pour rendre valables les résultats asymptotiques bien connus de la théorie de l'estimation.(On reviendra au n° 3 sur les autres cas).

On sait qu'alors l'estimation Z° du maximum de vraisemblance est sans biais appréciable, donc sa variance est pratiquement une perte d'information; et dans le cas régulier de Cramer (Réf.) $1/n$ H (H défini au Ch.IV, II, § 9) est le minimum (asymptotiquement équivalent à $\mathcal{V} Z^\circ$) de toutes les variances d'estimateurs sans biais possibles $\mathcal{V} Z$ construits à partir de la loi de distribution.

Tout ce qui est dit (en A et B) sur Z° (1ère interprétation géométrique notamment) est valable; à ceci près qu'il ne s'agit plus d'une multiplicité linéaire Z; que la corrélation $\sqrt{e}$ entre Z et Z° remplace $\sqrt{1-\rho^2}$ (où ρ désignait ρ_{xy}); que Z°, Z, $(Z - Z_0)$ ont des lois-limite de Laplace-Gauss, ce qui entraîne

$$\mathcal{E}(Z - Z_0)^2 = \mathcal{V} Z - \mathcal{V} Z^\circ$$

En particulier on peut avec Fortet (Réf.) distinguer trois cas :

1) Il existe pour ζ un estimateur sans biais

$$\bar{Z} = S\, z_i/n.$$

2) Il existe pour une fonction $\tau(\zeta)$ un estimateur sans biais

$$\bar{T} = S\, t_i/n.$$

3) Le cas général régulier, où existe l'estimateur Z° (faute de mieux).

Remarque : Le cas non régulier.

Dans une communication à l'Institut International de Statistique (1957), Alan Stuart signale des travaux récents(1) (1946 à 1952) concernant le cas non régulier, c'est-à-dire celui où, les conditions de régularité du 3ème cas n'étant pas remplies, l'estimation Z° du maximum de vraisemblance cesse d'exister ou perd ses droits. Il existe alors des cas où (l'estimation Z° n'existant pas) la variance de toute estimation ne peut descendre en-dessous d'une certaine limite qui est plus grande que $1/n$ H et qui peut jouer un rôle analogue à $\mathcal{V} Z^\circ$.

3/ - Cas où l'échantillon est de taille médiocre.

Les cas (1) et (2) ci-dessus restent inchangés. Mais au cas (3) $\mathcal{V} Z^\circ$ n'est plus une perte d'information, Z° étant biaisé. On peut lui

----

(1) Bhattacharyya — Sankya, 8. 1. 1946.
Chapman & Robbins — Ann. Math. Stat. 22(581) 1951.
Kiefer — Ann. Math. Stat. 23(627) 1952.

faire correspondre entre autres la perte $\lambda H/n$ et plus généralement $\mathcal{E} W\ [Z]$ avec [cf.Ch. IV § II 7]

$$W = \Phi\ [Z] - \Phi\ [\zeta] + \Sigma \lambda_j (a_j - A_j)$$

en posant $\mathcal{E}\ y_i = \sum_0^\infty a_j \zeta^j = 0$, $\quad S\ y_i / n \equiv \sum_0^\infty A_j Z^j = 0$

(On écrit Z au lieu de Z°, car il n'est plus vrai que p(Z) soit minimum).

Posant de même $\quad \Phi\ [\zeta] = f\ (a_o, a_1, a_2, \dots) = f$

$$\Phi\ [Z] = f\ (A_o, A_1, A_2, \dots) = f^*$$

et supposant la fonction f dérivable, on aurait

$$\frac{\partial W}{\partial A_j} = \frac{\partial f^*}{\partial A_j} - \lambda_j$$

qui, pour $Z = \zeta$ se réduit à $\lambda_j = \partial f / \partial a_j$

d'où

$$W = f^* - f + \Sigma \frac{\partial f}{\partial a_j} \cdot (a_j - A_j)$$

Toute fonction f convient alors, pourvu toutefois que la condition de convexité soit remplie :

$$\sum_j \sum_h \frac{\partial^2 f}{\partial a_j \partial a_h} (A_j - a_j)(A_h - a_h) \quad \text{définie négative.}$$

Toutefois il est exclu qu'on ait $\frac{\partial f}{\partial a_j} = \zeta^j$: W serait identique à 0.

Plus intéressant est le cas où f est elle-même une forme quadratique

$$f = \Sigma \Sigma\ b_{jh}\, a_j\, a_h$$

définie négative. D'où :

$$W \equiv \Sigma \Sigma\ b_{jh} (A_h - a_h) A_j$$

On notera que $\quad \overline{Y}(\mathcal{E}\ y_i) = \Sigma \Sigma\ \zeta^{j+h}\ A_j\, a_h \quad (\equiv 0)$

$$\overline{Y}^2 = \Sigma \Sigma\ \zeta^{j+h}\ A_j\, A_h$$

d'où $\quad \overline{Y}^2 = \Sigma \Sigma\ \zeta^{j+h}\ (A_h - a_h) A_j$

Lorsque les b_{jh} sont ζ^{j+h}, on retrouve donc

$$\mathcal{E}\, W \equiv H/n,$$

qui correspond à

$$\frac{\partial f}{\partial a_j} \equiv 0$$

(Voir Ch. IV, § II, 9).

Il serait beaucoup plus utile de connaître une expression de perte qui (pour n infini) serait asymptotiquement équivalente à 1/H n (comme l'est la variance de Z). Nous ne connaissons que

$$W = \bar{Y}^2/H^2$$

Des recherches ultérieures seraient nécessaires.

II - LE CHANGEMENT D'ECHANTILLON (à taille ou à coût constant) -

A - GENERALITES.

Après l'étude des suites de plans de sondage (d'une même famille) à effectifs décroissants, passons à la comparaison de plans de sondage (π), (π') appartenant à des familles distinctes.

Ces comparaisons n'ont de sens qu'à effectifs constants(1) ou à coûts constants.

Avec la méthode classique des sondages, la variance d'échantillonnage est employée constamment pour comparer la qualité de 2 (et plus) estimations du même paramètre ζ ; il s'agira :

- d'un même estimateur, employé avec des méthodes différentes de tirage de l'échantillon; par exemple avec une urne unique, on comparera les variances relatives aux tirages avec et sans remise des boules tirées dans l'urne;

- d'estimateurs différents, appliqués (sans changer la structure de sondage) à des échantillons tirés au sort par des procédés différents; par exemple avec probabilités égales d'une part, probabilités inégales de l'autre;

----

(1) A supposer que ceci ait encore une signification.

- voire d'estimateurs différents ou non, appliqués à des structures de sondage différentes; par exemple quand on modifie le nombre de strates, les caractères stratificateurs, le contenu des unités de sondage, ... bref le "découpage" de l'univers sondé.

Les problèmes du premier type sont d'ailleurs assez particuliers, l'absence de biais

$$\mathcal{E} Z = \mathcal{E} Z' = \zeta$$

étant postulée.

Il serait commode (et satisfaisant pour l'esprit) de pouvoir parler du gain d'information relatif à l'emploi de Z au lieu de Z' pour désigner la différence (positive) $\mathcal{E}(Z' - Z)^2$.

Mais il n'y a en général aucune raison a priori pour qu'on ait le droit de parler ainsi, c'est-à-dire de traiter cet écart positif comme la variance :

$$\mathcal{E}(Z' - Z)^2$$

Bien entendu, il est exclu que les deux échantillons à comparer soient indépendants, auquel cas on aurait :

$$\mathcal{E}(Z' - Z)^2 = \mathcal{V} Z' + \mathcal{V} Z$$

au lieu de :

$$\mathcal{E}(Z' - Z)^2 = \mathcal{V} Z' - \mathcal{V} Z$$

En fait quand on écrit cette dernière relation, autrement dit :

$$\mathcal{E}[(Z' - Z)(Z - \zeta)] = 0 \qquad (1)$$

on suppose l'existence d'un champ de probabilité sur lequel sont définis simultanément Z et Z', étant entendu que la loi de probabilité de Z (quel que soit Z') et celle de Z' (quel que soit Z) sont imposées. Le symbole $\mathcal{E}$ se trouvera du même coup défini.

Que signifie exactement (1), condition d'orthogonalité de $\overline{\zeta Z}$ et $\overline{ZZ'}$?

Sauf dans le cas de distributions particulières, il faut considérer que -, une fois l'échantillon tiré avec le plan de sondage π -, les $(Z - \zeta)$ sont des constantes quelconques et qu'il est nécessaire (et suffisant) d'avoir :

$$E(Z') = Z$$

où E désigne l'espérance mathématique lorsqu'est connu le grand échantillon (d'où on déduit la valeur Z).

On peut interpréter également (1) comme :

$$\mathcal{E}(Z' \mid Z) = Z$$

c'est-à-dire espérance mathématique liée par la connaissance de la valeur de Z; ces points de vue ne sont pas tout à fait équivalents : quand on entre dans le domaine des grands échantillons (population très nombreuse elle-même, les tirages pouvant en outre être avec remise), il est bien naturel qu'on puisse retrouver plusieurs fois la même valeur de Z. Enfin Z et Z' peuvent avoir une loi-limite, l'échantillon devenant infiniment grand; auquel cas la formule est forcément :

$$\mathcal{E}(Z' \mid Z) = Z$$

A supposer l'existence de deux lois-limites de Laplace-Gauss, avec $\mathcal{V}Z' > \mathcal{V}Z$, l'ellipse de probabilité de la loi de (Z Z') doit admettre la première bissectrice comme diamètre conjugué de la direction verticale.

En dehors de l'hypothèse ($n\ \infty$), la distribution commune $(Z\ Z')_n$ doit admettre comme courbe de régression de Z' en Z la droite d'équation Z' = Z; autrement dit, on doit avoir

$$Z' = Z + \vartheta, \qquad \text{avec} \qquad \mathcal{E}\ (\vartheta\ .\ Z) = 0$$

(et $\mathcal{E}\ \vartheta = 0$, qui est vérifiée si Z et Z' sont tous deux estimateurs sans biais de ζ).

Sauf si n est grand, (E Z' = Z) a la signification suivante :

Partant d'un échantillon du plan π (d'où une valeur de Z), on lui fait correspondre un ensemble d'échantillons du plan π', d'où un ensemble de valeurs Z'. On s'arrange pour avoir :

$$E\ Z' = Z$$

On s'arrange pour que, si l'on opère ainsi à partir de tous les échantillons équiprobables de π, on obtienne (un même nombre de fois) tous les échantillons équiprobables de π'.

D'ailleurs il suffit qu'une telle opération soit possible (sans aller jusqu'à l'exécution) pour qu'on ait le droit de convenir que

$\mathcal{V}Z' - \mathcal{V}Z = \mathcal{E}(Z - Z')^2$. On va rencontrer des cas où l'opération n'est pas possible.

B - 1ère INTERPRETATION GEOMETRIQUE.

Z et Z' étant représentés chacun dans un espace auxiliaire, avec

$$\mathcal{V}Z = \overline{OM}^2 \qquad \text{et} \qquad \mathcal{V}Z' = \overline{OM'}^2$$

on pourra avoir :

$$\overline{MM'}^2 = \overline{OM'}^2 - \overline{OM}^2$$

si la représentation de Z' est projetable orthogonalement sur celle de Z.

En particulier s'il s'agit de représentations à une seule dimension, il est nécessaire que les divisions découpées par Z et Z'(sur leurs axes respectifs) soient semblables.

1/ - Quelques cas de représentation à une dimension (estimation d'une moyenne).

Deux divisions d'abscisses $\sigma^2(\nu - n/n(\nu - 1)$ et $\sigma'^2(\nu - n/n(\nu - 1)$ sont semblables et même homothétiques. Sont semblables également :

$$\frac{\nu\sigma^2}{\nu - 1}\left(\frac{1}{n} - \frac{1}{\nu}\right) \qquad \& \qquad \frac{\nu'\sigma'^2}{\nu' - 1}\left(\frac{1}{n} - \frac{1}{\nu'}\right)$$

(y compris pour $\nu'\infty$: sondages exhaustif et bernoullien)

$$\frac{k\nu\sigma^2}{k\nu - 1}\left(\frac{1}{kn} - \frac{1}{k\nu}\right) \qquad \& \qquad \frac{\nu\sigma'^2}{\nu - 1}\left(\frac{1}{n} - \frac{1}{\nu}\right)$$

(sondages en grappes de même taille k).

etc.

2/ - Exemple : Tirage exhaustif et tirage bernoullien, avec n = 2.

Echantillon exhaustif	a, b	$\overline{X} = (a + b)/2$
Echantillon bernoullien	(a, b	$\overline{X}' = (a + b)/2$
	(a, a	$\overline{X}' = (a + a)/2 = a$

Il y a une probabilité $\nu - 1/\nu$ que la seconde boule tirée diffère de la première. Il vient :

$$\mathcal{E}\,(\overline{X}' - \overline{X})^2 = \frac{\nu - 1}{\nu}\,.\,0 + \frac{1}{\nu}\left(\frac{b - a}{2}\right)^2$$

Mais tous les couples (a, b) sont également possibles; d'où :

$$\mathcal{E}(\bar{X}' - \bar{X})^2 = \frac{1}{\nu} \frac{\Sigma(b-a)^2}{2\nu(\nu+1)} = \frac{\sigma^2}{2(\nu-1)} = \frac{\sigma^2}{2}\left[1 - \frac{\nu-2}{\nu-1}\right]$$

$$= \mathcal{V}\bar{X}' - \mathcal{V}\bar{X}$$

3/ - Un cas de représentation à plusieurs dimensions (estimation d'une moyenne) :

Tirages exhaustif et bernoullien dans un sondage stratifié.

Si dans chaque strate l'échantillon bernoullien est déduit - d'un échantillon exhaustif donné - par un procédé aléatoire convenable, on a :

$$\mathcal{E}(\bar{X}_h' - \bar{X}_h)^2 = \mathcal{E}(\bar{X}_h' - \bar{x}_h)^2 - \mathcal{E}(\bar{X}_h - \bar{x}_h)^2$$

et, en pondérant par $(\nu_h/\nu)^2$ et totalisant, - en admettant en outre l'indépendance des opérations de chaque strate par rapport aux autres - , il vient :

$$\mathcal{E}(\bar{X}' - \bar{X})^2 = \mathcal{E}(\bar{X}' - \bar{x})^2 - \mathcal{E}(\bar{X} - \bar{x})^2$$

C - 2ème INTERPRETATION GEOMETRIQUE.

Z et Z' sont représentés dans un espace euclidien auxiliaire , avec

$$\mathcal{V}Z = \overline{OM}, \qquad \mathcal{V}Z' = \overline{OM'}$$

Quand les pertes d'information sont de la forme :

$$\mathcal{V}Z = \alpha + \beta + \gamma, \qquad \mathcal{V}Z' = \alpha' + \beta' + \gamma'$$

il vient :

$$\mathcal{E}Z' - \mathcal{V}Z = (\alpha' - \alpha) + (\beta' - \beta) + (\gamma' - \gamma)$$

Il est clair que, si α' est fonction linéaire de α, de même β' et β, γ' et γ, - alors $\mathcal{V}Z' - \mathcal{V}Z$ est aussi une perte d'information, et c'est $\mathcal{E}(Z' - Z)^2$.

Exemple : Sondages à 2 degrés. Comparaison des tirages exhaustif et bernoullien. Cas particulier où $n_h = \bar{n}$; $\nu_h = \bar{\nu}$.

α' est proportionnel à $1/m$, et α à $(1/m - 1/\mu)$

β' est proportionnel à $1/n$, et β à $(1/n - 1/\nu)$
γ' est proportionnel à $1/mn$, et γ à $(1/m - 1/\mu)(1/n - 1/\nu)$

On est donc bien dans le cas où $\mathcal{V}\bar{X}' - \mathcal{V}\bar{X}$ est une variance.

Remarque.

Il est clair que, si le mode de tirage différait pour le 2ème degré seul, ou pour le 1er degré seul, il suffirait de déduire l'échantillon bernoullien de l'échantillon exhaustif par le procédé aléatoire.

Plus généralement, partant d'une sous-population de $(m \times \bar{n})$ unités distinctes (dont la moyenne est $\bar{X}$), on éliminera (par tirages au sort équiprobables) un nombre quelconque λ d'unités qu'on remplacera (par tirages au sort équiprobables) par λ unités figurant déjà dans l'urne et qui y figureront ainsi deux fois), etc

Ainsi il est possible de construire un échantillon bernoullien à partir de l'échantillon exhaustif supposé donné,et de telle façon que :

$$\mathcal{E}(\bar{X}' - \bar{X})^2 = \mathcal{V}\bar{X}' - \mathcal{V}\bar{X}.$$

D - LE CAS GENERAL.

1/ - Il n'y a pas de distance entre deux estimateurs.

La comparaison des tirages bernoulliens et exhaustifs constitue en fait un cas d'exception; car lorsqu'on passe en revue les diverses circonstantes courantes où la qualité d'estimateurs Z et Z' se juge en étudiant le signe de $\mathcal{V}Z - \mathcal{V}Z'$, on est forcé de constater que cette différence n'est ni $\pm\mathcal{E}(Z - Z')^2$, ni tout autre distance entre Z et Z', ou entre les plans de sondage (π) et (π'). En effet (tout en ayant : $\mathcal{E}Z' = \mathcal{E}Z = \zeta$) : (π') est conçu de façon : soit à avoir $\mathcal{V}Z - \mathcal{V}Z' \geqslant 0$ (en tout état de cause, ou bien dans la majorité des cas); soit à réduire le coût d'enquête, avec une certitude ou des chances sérieuses d'avoir alors : $\mathcal{V}Z - \mathcal{V}Z' < 0$.

Dans le premier cas, on s'arrange pour que les valeurs de Z les plus éloignées de ζ ne soient pas (ou soient très rarement) prises par Z'; pour cela certains échantillons E du plan (π) sont impossibles avec le plan (π'). Dans le second cas, on arrive au même fait pour une simple raison de commodité.

Ainsi à un certain champ de valeurs de Z ne correspond aucune valeur de Z'; et parler de

$$\mathcal{V}(Z' - Z)^2 = \mathcal{V}Z + \mathcal{V}Z' - 2\,\mathrm{Cov}(Z\,Z')$$

ne signifie rien, car Cov(Z Z') n'est pas défini; alors parler de la distance (Z Z') risque fort de ne rien vouloir dire. On vise ainsi les cas suivants :

Comparaison d'un sondage systématique ou en grappe et du sondage ordinaire;

Comparaison d'un sondage stratifié et du sondage ordinaire;

Comparaison d'un sondage "équilibré" ("balanced sampling") et du sondage ordinaire;

Comparaison des tirages avec probabilités égales et probabilités inégales, etc.

Ajoutons que, si l'on compare par exemple deux découpages distincts des unités de sondage, ou deux découpages des strates, il peut ne plus même exister de champ où Z et Z' soient définies simultanément.

2/ - Que calcule-t-on en fait ?

Cependant la technique des sondages nous apprend, connaissant l'échantillon E' de (π'), à estimer non seulement $\mathcal{V}Z'$, mais aussi $\mathcal{V}Z$ et $\mathcal{V}Z - \mathcal{V}Z'$[1]. On pourrait d'ailleurs estimer $\mathcal{V}Z$ et $\mathcal{V}Z'$ séparément avec deux enquêtes distinctes, avec respectivement un échantillon E de (π) et E' de (π'); mais il serait à craindre que l'erreur d'échantillonnage sur $\mathcal{V}Z - \mathcal{V}Z'$ (estimations indépendantes) soit plus grande que ($\mathcal{V}Z - \mathcal{V}Z'$) elle-même; en employant un seul échantillon (soit E') on ne court guère de risque de se tromper sur le signe même de $\mathcal{V}Z - \mathcal{V}Z'$.

Faisons quelques remarques :

a) On estime Z' (et non Z) à l'aide de E'.

b) On estime $\mathcal{V}Z$ et $\mathcal{V}Z'$ à l'aide de E'; mais on ne saurait estimer $\mathcal{V}Z'$ à l'aide de E (échantillon de π). Donc E' permet de calculer une sorte de distance entre le plan de sondage (π') et le plan de référence (π), et non le contraire.

c) On sait (souvent)[2] estimer sans biais $\mathcal{V}Z$, mais il ne faut pas s'illusionner sur la signification de ces mots :

---

(1) Ceci fut pour nous un objet de profond étonnement à la première lecture du rapport de R. Jessen sur les sondages agricoles en IOWA (Réf. T1).

(2) Dans les cas les plus simples, l'analyse de variance fournit un mécanisme de calculs correct (du moins en première approximation).

Si l'on tirait (du même plan (π') un nombre infini d'échantillons E' indépendants, on reconstituerait la distribution théorique de Z' et celle de $\mathcal{V}Z'$, avec la répartition statistique des valeurs des estimateurs. En revanche la distribution des valeurs calculées pour estimer $\mathcal{V}Z$, tout en admettant $\mathcal{V}Z$ pour espérance mathématique, ne coïnciderait pas du tout avec la distribution théorique de $\mathcal{V}Z$ (telle qu'on l'obtiendrait au contraire avec une infinité d'échantillons indépendants E du plan π).

d) On peut toujours écrire $\mathcal{E}(Z_1 - \zeta)^2 - \mathcal{E}(Z' - \zeta)^2 = \mathcal{E}(Z_1 - Z')^2$ avec $\mathcal{E}(Z_1 - Z')(Z' - \zeta) = 0$, c'est-à-dire $\mathcal{E}'[Z_1 - Z' \mid Z'] = 0$, où $\mathcal{E}'$ désigne l'espérance mathématique dans le champ de probabilités de Z'; les valeurs de Z_1 (pour un échantillon donné de π') admettent Z' pour moyenne arithmétique; puis on s'arrange pour avoir $\mathcal{E}'(Z_1 - \zeta)^2 \equiv \mathcal{V}Z$, sans pour autant que Z_1 et Z aient la même distribution.

On voit, en résumé, que (même en s'en tenant aux estimations sans biais) les écarts entre pertes d'information se prêtent mal à une interprétation simpliste comme distance entre deux points.

CHAPITRE VIII

SUR DIVERS CONCEPTS D'INFORMATION APPLICABLES AUX SONDAGES

I - LA QUANTITE D'INFORMATION DE R.A. FISHER -

Alors que nous voyons dans $1/n$ H la plus petite perte d'information qu'on puisse obtenir (à condition d'employer l'estimateur adéquat), R.A. Fisher appelle n H l'information contenue dans l'échantillon, c'est-à-dire le plus grand gain d'information que puisse fournir le même échantillon (du moins si l'on est bien dans le "cas régulier") (Réf. Fi). On suppose toujours n très grand.

Dans le cas de l'estimation simultanée sans biais de deux paramètres ζ_1, ζ_2 d'une loi de distribution, la perte d'information $\mathcal{V}(u_1Z_1 + u_2Z_2)$ est une forme quadratique; égalée à W^2 c'est l'équation tangentielle d'une ellipse; lorsque n est assez grand, Dugué a établi l'existence de l'ellipse d'information intérieure à la précédente quels que soient Z_1 et Z_2; son équation est $Q = W^2$; la matrice de la forme quadratique Q est ce que Fisher a défini comme étant l'information.

Tels sont les rapports entre la perte d'information et l'information de Fisher.

II - EXTENSION AUX SONDAGES BERNOULLIENS -

Pour une loi de Laplace, l'estimateur efficace de $\bar{x} = \mu$ est la moyenne $\bar{X}$, la variance minimum des estimateurs est σ^2/n et l'information au sens de Fisher est n/σ^2. Ceci s'étend à toutes les lois qui admettent un résumé exhaustif d'ordre 1.

L'idée la plus naturelle est donc de chercher à faire une théorie où l'information serait $1/\mathcal{V}\bar{X}$, c'est-à-dire n/σ^2 quelle que soit la loi de distribution. (n étant quelconque cette fois).

On serait même tenté d'appeler quantité d'information l'inverse $(1/\mathcal{V}Z)$ de la variance, quel que soit l'estimateur Z.

Si l'on effectue les tirages bernoulliens dans une urne comprenant un nombre fini de boules, la définition précédente serait compatible avec la formule de récurrence (de Schutzenberger). Faisons deux tirages successifs d'une boule dans une urne renfermant quatre boules; la quantité d'information apportée serait :

$$\frac{1}{\sigma^2} + \frac{1}{4} \sum_{1}^{4} \frac{1}{\sigma^2} = \frac{2}{\sigma^2}$$

et pour n tirages

$$n/\sigma^2 = 1/\mathcal{V}\bar{X}$$

III - CAS DES SONDAGES EXHAUSTIFS -

En revanche avec des tirages exhaustifs, la même formule donne, pour deux tirages :

$$\frac{1}{\sigma^2} + \mathcal{E}\left(\frac{1}{s_i^2}\right)$$

où s_i^2 désigne la variance des (ν - 1) boules restantes après le premier tirage.

De l'identité

$$\mathcal{E}\left(\frac{\nu - 1}{\nu - 2}\; s_i^2\right) = \frac{\nu}{\nu - 1}\,\sigma^2$$

on ne peut tirer la valeur de $\mathcal{E}\;(1/s_i^2)$ mais on peut affirmer en tous cas qu'on n'a pas en général :

$$\mathcal{E}\left(\frac{1}{s_i^2}\right) = \frac{\nu}{\nu - 2}\;\frac{1}{\sigma^2}$$

et par conséquent qu'on n'a pas :

$$\frac{1}{\sigma^2} + \mathcal{E}\left(\frac{1}{s_i^2}\right) = \frac{1}{\sigma^2}\left(1 + \frac{\nu}{\nu - 2}\right) = \frac{2}{\sigma^2}\;\frac{\nu - 1}{\nu - 2} = \frac{1}{\mathcal{V}\,\bar{X}}$$

En résumé : si l'on s'en tenait aux tirages indépendants, on pourrait accepter de transposer l'information de Fisher en théorie des sondages, de telle sorte que l'inverse de la variance de l'estimateur mesure de l'information apportée par celui-ci.

Mais on ne peut étendre valablement cette dernière convention si l'on passe du tirage bernoullien au tirage exhaustif des échantillons.

IV - DISCUSSION -

a) D'ailleurs il n'est pas rationnel de partir du sondage bernoullien et d'essayer de remonter au sondage exhaustif. Ce qu'il convient de faire, c'est au contraire de poser pour le sondage exhaustif des définitions cohérentes, puis d'en déduire celles du sondage bernoullien par passage à la limite lorsque l'effectif de la population tend vers l'infini.

b) En outre (mais ceci n'est pas un argument déterminant) le concept d'information ne correspond pas, quand on se place au point de vue des sondages, à celui posé par Fisher. Si je possède déjà un échantillon de 3 000 unités de sondage par exemple, et si j'en prends 100 de plus, il ne paraît guère possible d'admettre que ces 100 unités m'apportent autant d'information que les 100 premières unités qu'on a tirées de la population (ceci abstraction faite de toute considération de coût). C'est pourquoi, même lorsqu'on tire l'échantillon à la manière de Bernoulli, la définition de l'information inspirée par Fisher ne nous paraît pas finalement à retenir.

A cet égard Lindley (dont le concept d'information diffère du nôtre) a adopté la même position. L'information "apportée" par une unité est pour lui fonction concave du rang de tirage.

Ceci n'empêchera d'ailleurs pas de continuer, comme par le passé, à utiliser l'inverse du rapport des variances pour juger de la précision relative - ou de l'efficacité[1] comme on dit - de deux estimateurs, du moment qu'il s'agit de précision ou d'efficacité, - non de l'information.

V - SOLUTION -

On peut chercher à bâtir d'abord une théorie de l'information pour les sondages exhaustifs dans une urne unique. Cette information doit - avant tout - satisfaire à la condition générale de récurrence. Or le Lemme

----

(1) A noter que Fisher (1922) appelle efficacité, non le rapport des inverses de variance mais le rapport des effectifs d'échantillon nécessaires pour obtenir la même variance; définitions qui ne coïncident que si la variance est de la forme a/n (Voir, Stuart. Réf. citée).

$$\mathcal{E}\left(\frac{n}{n-1}\ s_n^2\right) = \frac{\nu}{\nu - 1}\ \sigma^2$$

a semblé particulièrement commode pour l'application de la formule de récurrence; et ceci a conduit à rechercher à repérer le niveau d'information par une fonction de la forme :

$$t(\nu,\ n) = k(\nu,\ n)\,\sigma^2$$

l'effectif de la population étant ν, celui de l'échantillon étant n, σ désignant l'écart-type de la variable x dans la population sondée.

On appellera gain ou perte d'information l'écart entre deux niveaux. Ce niveau d'information doit s'élever lorsque n va de 1 à $(\nu - 1)$.

Avec un échantillon de taille nulle, on n'a pas d'information du tout.

Avec un échantillon de taille ν, l'information est complète, totale.

Le problème essentiel est de savoir choisir pour ces deux cas extrêmes des repères convenables. Il est naturel de chercher à repérer par 0 l'absence d'information, par A ou l'infini l'information complète. On dira alors qu'on repère ou mesure l'information apportée par l'échantillon.

On va montrer que ce point de vue est indéfendable pour les enquêtes par sondage, avec $t(\nu,\ n)$. Au contraire, rien ne s'oppose à ce qu'on repère par $-\infty$ l'absence d'information. On dira alors qu'on repère ou mesure l'information perdue par le sondage.

VI - IMPOSSIBILITE DE MESURER L'INFORMATION APPORTEE PAR UN SONDAGE AVEC $t(\nu,\ n)$ -

Si le choix de l'information reste largement arbitraire quand on n'envisage que le sondage exhaustif simple, il ne faut pas oublier qu'on se propose d'aboutir à une théorie applicable à tous les plans de sondage communément employés par les techniciens.

Une autre difficulté est de savoir comment vont se combiner les informations fournies par chacune des urnes sondées (strates, sous-strates, unités non élémentaires). Cette combinaison est-elle de même nature qu'une addition (déduction faite des doubles emplois), ou bien qu'un produit ? Mais faire le produit revient à ajouter les logarithmes et tout dépend finalement d'une convention de langage : va-t-on appeler u ou log u l'information ?

a) Sondage à deux degrés.

Considérons un sondage à deux degrés, avec le cas limite du sondage en grappe (tirages exhaustifs avec probabilités égales) : si on possède une information partielle sur certaines unités primaires et une information complète sur certaines autres (grappes entières), on n'a pas une information complète sur la population.

Par conséquent, que les informations apportées par chaque unité primaire s'ajoutent ou qu'elles se multiplient les unes par les autres, - on peut dire a priori que l'information complète ne doit pas être mesurée par $+\infty$; elle doit être exprimée par un nombre fini. On va le supposer par la suite(1).

b) Addition des informations.

Lorsqu'on effectue des sondages indépendants dans une même population, il semble a priori que les informations apportées devraient s'ajouter - du moins d'après les idées communément admises sur l'information.

De sorte que, lorsqu'on ajoute l'information d'un sondage à celle d'un recensement, on devrait retrouver l'information du seul recensement (vu que le sondage n'apporte rien de plus). Ainsi, logiquement, l'information complète (celle procurée par le recensement) devrait s'exprimer par $+\infty$. (On observera qu'il en est de même si l'addition porte sur les logarithmes des informations et non sur les informations elles-mêmes).

En réalité ce second point de vue est logique sans être tout à fait décisif. On peut convenir en effet de limiter les recherches à un seul sondage simple exhaustif, - ce qui exclut toute possibilité d'ajouter à un sondage à 100% un autre sondage exhaustif à 5%. Lorsque l'urne est vide, l'opération s'arrête.

Ainsi il n'y a peut-être pas de contradiction flagrante entre (1) et (2), mais il y a tout de même là une restriction notable, une option inévitable entre deux concepts.

c) Considérons à présent un sondage stratifié.

S'il est une strate sur laquelle on ne possède aucune information, il est conforme aux habitudes de pensée des "sondeurs" de dire qu'on ne possède aucune information sur la population entière (mais seule-

----

(1) Tel n'est pas le cas pour l'information de Lindley (voir plus loin §IX).

ment sur les autres strates); c'est d'ailleurs un point sur lequel le statisticien dépourvu de formation sondage a une opinion différente.

Une conséquence fâcheuse en découle si l'on convient de repérer l'absence d'information par le niveau 0 (et non par le niveau $-\infty$) :

Si l'on veut conserver cette vue des techniciens, il ne faut pas que les informations apportées sur chaque strate s'ajoutent ou se combinent linéairement pour donner l'information relative à la population; tandis qu'il n'est pas exclu que les informations de strate se multiplient les unes par les autres (donc que leurs logarithmes s'additionnent).

Or il est facile de voir que c'est là une exigence incompatible avec notre hypothèse de travail : $t(\nu, n) = k(\nu, n)\sigma^2$ valable à l'intérieur des strates.

Si l'on pose (avec des poids convenables ϖ)

$$\mathcal{L}\, t = \sum_h \varpi_h(\nu_h, n_h)\, \mathcal{L}\, t(\nu_h, n_h)$$

ceci implique (quand on épuise toutes les strates) :

$$\max \mathcal{L}\, t = \sum_h \varpi_h(\nu_h, n_h)\left[\mathcal{L}\, k(\nu_h, n_h) + \mathcal{L}\, \sigma_h^2\right]$$

Considérons alors un sondage à deux degrés. Si toutes les unités primaires sont sondées, le sondage à deux degrés devient sondage stratifié, la perte a la forme ci-dessus; si au contraire toutes les unités secondaires sont sondées dans l'échantillon d'unités primaires, on a un sondage en grappes dont la perte d'information est de la forme :

$$\mathcal{L}[k(\mu, m)\, \sigma_0^2] \qquad \text{avec} \qquad \sigma_0^2 = \frac{1}{\mu} \Sigma (\overline{x}_h - \overline{x})^2$$

et quand l'échantillon recouvre toute la population

$$\max \mathcal{L}\, t = \mathcal{L}\, k(\mu, \mu) + \mathcal{L}\, \sigma_0^2$$

Comme on a supposé que l'information restait bornée si la population entière est tirée, les deux expressions de max $\mathcal{L}\, t$ sont incompatibles en général (puisqu'elles imposent une condition aux paramètres de la distribution).

d) Un autre fait est qu'on devrait pouvoir (toutes choses égales d'ailleurs) appeler sondage optimal celui qui apporte le maximum d'information. La théorie classique connait déjà les répartitions d'échantillon optimales au sens de Neyman (telles que la variance

d'échantillonnage et le coût du sondage soient l'un minimum et l'autre constant). Ainsi il est souhaitable que ces deux points de vue soient conciliables (sans coïncider pour autant).

Admettons que le coût $\mathcal{C}$ soit combinaison linéaire des effectifs échantillon de strate.

Si l'on convenait que les informations de strate s'ajoutent, un cas particulièrement gênant serait celui où l'information de strate serait de la forme :

$$\lambda_h \quad n_h$$

puisque le maximum d'information (pour un coût total $\mathcal{C} = \sum_h c_h n_h$ donné d'avance) s'obtiendrait en concentrant l'échantillon sur les strates dont le coefficient λ_h / c_h est le plus grand (contrairement à la technique admise.

Or on trouverait effectivement une expression de ce type pour l'information apportée dans des cas très simples. Posons :

$$t(\nu, n) = \frac{\nu \sigma^2}{\nu - 1} f(\nu, n)$$

Reprenons une idée de M. Fonsagrive (après M. Chartier, Réf. Chartier 2). On considère deux strates d'effectif ν formées par tirage au sort exhaustifs dans une urne d'effectif $2n$. On tire n boules de chaque strate.

Il est raisonnable d'admettre que l'information apportée est en moyenne la même que si (2n) boules étaient tirées de l'urne primitive. D'où une relation :

$$\mathcal{E}\left[\frac{\nu}{\nu - 1}(\sigma_1^2 + \sigma_2^2) f(\nu, n)\right] = \frac{2\nu}{2\nu - 1}\sigma^2 f(2\nu, 2n)$$

Grâce au lemme fondamental, on a :

$$\mathcal{E}\sigma_1^2 = \mathcal{E}\sigma_2^2 = \frac{2(\nu - 1)}{2\nu - 1}\sigma^2$$

D'où : $$2 f(\nu, n) = f(2\nu, 2n)$$

cas particulier de l'équation fonctionnelle

$$\rho \ f(\nu, n) = f(\rho\nu, \rho n)$$

En partageant en deux l'urne primitive, on est parvenu à l'équation avec $\rho = 2$; mais en la partageant en ρ parties égales, on obtientrait la même équation pour les diverses valeurs entières de ρ. La solution de cette équation compatible avec les autres conditions est justement λn.

Il est donc établi qu'on ne peut mesurer l'information apportée par un sondage avec l'expression $t(\nu, n) = \sigma^2 k(\nu, n)$.

La notion de quantité d'information apportée ayant fait l'objet de travaux importants à la limite du domaine des sondages, on va en passer en revue quelques-uns.

VII - L'INFORMATION DE SHANNON -

Le vocable "information" employé d'abord par Fisher a été repris par Shannon pour l'étude de problèmes de transmission. Good(1) notamment a pu exprimer cette "opinion raisonnable" qu'il n'y a pas grand'chose de commun entre l'information de Shannon et celle de Fisher.

C'est Schutzenberger(2) qui, plus tard, a montré qu'elles entraient toutes deux dans le cadre d'une théorie générale de l'information.

Shannon appelait information d'un message la somme des informations des divers symboles qui le composent, - et information d'un symbole déterminé l'espérance mathématique ou plutôt la moyenne, changée de signe, du logarithme de la fréquence de son apparition.

Le problème pour Shannon était de faire occuper le moins de temps possible une ligne télégraphique par les messages; il fallait pour cela choisir pour les divers symboles possibles un code optimum tenant compte de la fréquence de chaque symbole.

La quantité d'information d'un message détermine finalement le temps pendant lequel on permet au message d'occuper la ligne quand on emploie ce code.

De leur côté, les statisticiens éprouvent le besoin de pouvoir définir une sorte de mesure (additive) de leurs travaux, en fonction de quoi ils répartiraient les moyens limités (personnel, matériel, crédits) dont ils disposent.

----

(1) Discussion à la suite de la communication de Barnard, Réf. Bar.

(2) Réf. Sch.

La "capacité" de la ligne a pour homologue en statistique la capacité du bureau; mais quel sera l'homologue de la quantité d'information ? Ceci dépend de la manière dont on va transposer en statistique les deux notions de message et de fréquence.

Cette transposition a été tentée à diverses reprises et notamment par Barnard[1].

VIII - LA QUANTITE D'INFORMATION DE BARNARD -

Barnard envisage une théorie suffisamment générale pour que le même symbole représente aussi bien un message (en théorie des transmissions), un problème (pour un calculateur électronique ou toute autre machine à calculer), ou une proposition (en statistique considérée comme un secteur de la logique).

Notons en passant que la notion de "quantité d'information" de Shannon a été effectivement adoptée par la cybernétique et est finalement revenue jusqu'au statisticien comme partie intégrante du "jargon" des spécialistes du matériel électronique.

Le "message" est donc transposé en une "proposition", et la "fréquence" devient une "probabilité", la probabilité que la proposition soit exacte. C'est ici que les difficultés surgissent.

Quel intérêt aurait-on à remplacer une probabilité p variant de 0 à 1, par une expression (-log p) variant de 0 à l'infini ?

Est-il plus intéressant d'avoir affaire à des expressions qui s'ajoutent, ou au contraire qui se multiplient les unes par les autres, quand on envisage simultanément plusieurs propositions disjointes ?

$$\log p + \log p' = \log pp'$$

Barnard arrive vite à la conclusion que -log p ne mérite pas d'être tenu pour la quantité d'information.

L'information serait une certaine grandeur (voire une certaine fonction) qu'on peut attacher à une proposition, au même titre qu'une probabilité; mais ce n'est certainement pas une simple fonction de la probabilité (ni -log p, ni une autre).

Intéressons-nous plus spécialement à l'information en matière d'estimation statistique :

---

(1) Réf. G. Barnard m'a prié (1960) de préciser qu'il avait exposé (1950) une façon de concevoir l'information, mais en admettait volontiers d'autres.

Etant donné une loi de distribution $f(x, \zeta)dx$, où le paramètre ζ est à estimer, soit X_i une valeur réellement observée pour x ; l'information apportée par cette observation pourrait être :

$$- \log f(X_i, \zeta)$$

autrement une certaine fonction de ζ, et même une fonction de ce que Fisher appelle la vraisemblance (likelihood).

Pour Fisher (et Barnard) il n'est ni probabilité a posteriori ni loi de Bayes; une probabilité est toujours a priori, avant qu'on ait fait les expériences, les observations; une fois franchi le pas, il n'existe plus de probabilités, mais des vraisemblances (et des informations).

Et Barnard précise que Shannon, s'il était à sa place, considèrerait ζ comme fixe, X_i comme variable, et prendrait la moyenne de (-log) sur l'ensemble des valeurs que X_i est susceptible de prendre.

IX - LA QUANTITE D'INFORMATION DE LINDLEY -

Quand on admet la loi de Bayes, et que par conséquent on ne conçoit pas de théorie de l'estimation sans une loi de probabilité a priori, on est conduit assez naturellement à définir la quantité d'information en fonction de la probabilité a posteriori, et non plus de la "vraisemblance".

Il est alors possible de conserver presqu'intégralement la définition de Shannon; on abandonne seulement le signe moins qui précède le logarithme, de façon à définir, pour chaque état de nos connaissances, un niveau d'information négatif; quand on gagne de l'information, on s'élève vers la cote zéro. Nous suivons ici le point de vue de Lindley, le plus récent (Réf. Li)(1).

Soit $p(\zeta)$ la probabilité a priori et $p(\zeta|x)$ la probabilité a posteriori de ζ sachant que x est le résultat de ou des observations.

Les niveaux d'informations successifs sont, pour des tirages bernoulliens :

$$I_0 = \int p(\zeta) \log p(\zeta) d\zeta$$

$$I_1 = \int p(\zeta|x) \log p(\zeta|x) d\zeta$$

$$\overline{I}_1 = \int I_1\, p(x) dx \qquad \text{(information moyenne)}$$

----

(1) Il ne diffère pas essentiellement de celui de ses prédécesseurs, McMillan, Blackwell, Bohnenblust-Shapley-Sherman, etc., de l'avis de Lindley.

où p(x) désigne la loi de probabilité a priori de l'observation x.

Se plaçant au point de vue "séquentiel", on définit le niveau I_2 d'information après deux résultats d'observation x, x', en substituant

$$p(\zeta \mid x, x') \qquad \text{à} \qquad p(\zeta \mid x)$$

dans la formule; on définit de même le niveau I_3 avec

$$p(\zeta \mid x, x', x''),$$

etc.

Le gain d'information apporté par les unités de sondage tirées aux rangs n+1, n + 2, ... m est égal à la différence du niveau $I_m - I_n$. Lindley montre que c'est une fonction "concave".

Comparaison avec la perte d'information.

Le caractère essentiel de la perte d'information est que le niveau I_∞ n'est pas rejeté à l'infini. Au contraire, dans des cas très simples, on constate que le niveau I_n de Lindley tend vers l'infini avec n; par exemple lorsqu'on a (exemple du § X)

$$I_n = \frac{1}{2} \log(\lambda^2 + n) + \text{constante}$$

Par ailleurs le niveau I_0 serait (pour la perte d'information) rejeté à l'infini; alors qu'ici on part toujours d'un niveau fini. Ceci tient au fait qu'on ne part jamais d'une absence complète d'information, on se donne toujours a priori une distribution (qui peut en fait provenir d'un sondage antérieur).

X - ETUDE D'UN EXEMPLE DE LINDLEY -

Etudions avec Lindley une population infinie, distribuée suivant une loi de Laplace-Gauss de moyenne aléatoire et d'écart-type σ.

Considérons a priori ζ comme une autre variable de Laplace-Gauss, de moyenne μ (constante celle-la) et d'écart-type τ.

Supposons d'abord l'échantillon d'effectif n = 1, la probabilité élémentaire du couple aléatoire (x, ζ) est

$$(2\pi)^{-1} (\sigma\tau)^{-1} \exp \left\{ -\frac{1}{2} \left[\frac{(x-\zeta)^2}{\sigma^2} + \frac{(\zeta-\mu)^2}{\tau^2} \right] \right\} dx d\zeta$$

et l'exposant de l'exponentielle peut aussi s'écrire :

$$-\frac{1}{2}\left[\frac{(x-\mu)^2}{\tau^2+\sigma^2}+\frac{\tau^2+\sigma^2}{\tau^2\sigma^2}\left(\zeta-\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}\right)^2\right]$$

ce qu'on interprète en disant que la loi de probabilité a priori de x est de Laplace-Gauss, de moyenne μ et d'écart-type $\sqrt{\tau^2+\sigma^2}$; et que la loi de probabilité a posteriori de ζ est de même type, mais de moyenne : $\frac{\sigma^2\mu+\tau^2 x}{\sigma^2+\tau^2}$ et d'écart-type $\frac{\sigma\tau}{\sqrt{\sigma^2+\tau^2}}$. Il est utile ici de faire ressortir que :

THEOREME -

Si ζ a une distribution a posteriori de Laplace-Gauss, la quantité d'information est égale au logarithme (neperien) de l'inverse de son écart-type, diminué de $\log\sqrt{2\pi}$ (démonstration immédiate).

En conséquence, le gain d'information $I_1 - I_0$ est égal à :

$$\log\sqrt{\sigma^2+\tau^2}-\log\sigma=\frac{1}{2}\log\left(1+\frac{\tau^2}{\sigma^2}\right)$$

Passons de là à un échantillon de taille $n = 2$; il vient :

$$\frac{(x_1-\zeta)^2}{\sigma^2}+\frac{(x_2-\zeta)^2}{\sigma^2}+\frac{(\zeta-\mu)^2}{\tau^2}\equiv\frac{2\tau^2+\sigma^2}{\tau^2\sigma^2}\left(\zeta-\frac{\tau^2(x_1+x_2)+\sigma^2\mu}{2\tau^2+\sigma^2}\right)^2+\dots$$

et le gain d'information est :

$$\frac{1}{2}\log(\sigma^2+\tau^2/\sigma^2+2\tau^2)=I_2-I_1$$

Avec un échantillon de taille n, le gain d'information est :

par rapport au cas où $n = 0$

$$I_n-I_0=\frac{1}{2}\log\left(1+\frac{\lambda^2}{n}\right)+\log n \qquad \text{en posant } \lambda^2=\frac{\sigma^2}{\tau^2}$$

$$=\frac{1}{2}\log(\lambda^2+n)$$

par rapport à un échantillon de taille $n - 1$, il est :

$$I_n-I_{n-1}=\frac{1}{2}\log\left(1+\frac{\lambda^2}{n}\right)-\frac{1}{2}\log\left(1+\frac{\lambda^2}{n-1}\right)+\frac{1}{2}\log\left(\frac{n}{n-1}\right)$$

Quant à la loi a posteriori de ζ , c'est une loi de Laplace-Gauss

de moyenne $\dfrac{\sigma^2\mu + n\tau^2\overline{X}}{\sigma^2 + n\tau^2}$ et d'écart-type $\sqrt{\dfrac{\tau^2\sigma^2}{\sigma^2 + n\tau^2}}$

ou de moyenne $\dfrac{\lambda^2\mu + n\overline{X}}{\lambda^2 + n}$ et d'écart-type $\sqrt{\dfrac{\sigma^2}{\lambda^2 + n}}$

On peut comparer gain d'information et réduction de la perte d'information (variance), ou plutôt (si l'on fait abstraction de log(n/n - 1) et de $\frac{1}{2}$) comparer

$$\log\left(1 + \frac{\lambda^2}{n}\right) - \log\left(1 + \frac{\lambda^2}{n-1}\right) \quad \text{et} \quad \left(\frac{\sigma^2}{n} - \frac{\sigma^2}{n-1}\right)$$

Lorsque n est grand (en supposant τ^2 du même ordre que σ^2), ces deux expressions sont des infiniment petits équivalents (au facteur τ^2 près).

Quant à la loi a posteriori de ζ , elle correspond assez bien à nos habitudes de pensée : il suffit d'imaginer qu'un sondage antérieur nous a procuré une moyenne μ et une variance (estimée) τ^2 et qu'on les prend comme paramètres de loi de probabilité a priori; alors on retrouve comme moyenne de la loi a posteriori l'estimateur habituel.

La difficulté est qu'il faut donc toujours commencer par faire une première estimation sans probabilité a priori, - ce qu'on déguise en l'appelant : choix des paramètres μ et τ .

En dehors de ce cas simple, de peu d'intérêt pratique, l'information de Lindley (comme l'estimation par la méthode de Bayes) diverge franchement de la présente théorie.

XI - POINT DE VUE SUR CETTE INFORMATION -

Le problème de l'estimation statistique peut être abordé de trois façons différentes :

1) En se donnant la loi de répartition de X et la probabilité a priori du paramètre à estimer ζ : comme Lindley ;

2) En se donnant seulement la loi de répartition de X, - la grandeur à estimer étant une constante inconnue non aléatoire : c'est le point de vue de R.A. Fisher;

3) En ne précisant même pas la loi de répartition de X, loi qui en pratique n'est pas connue et ne présente que des rapports éloignés avec les lois théoriques usuelles ou non : c'est le point de vue de la théorie des sondages.

Le propre de la théorie des sondages est de faire des estimations en ignorant la forme des lois de distribution des variables en jeu.

A fortiori est-il indispensable d'ignorer les lois de probabilité a priori des ζ.

A - Cependant, dans le cas de populations sondées très grandes (et c'est un cas limite aussi important en pratique qu'en théorie) les X échantillon devenant indépendants, la théorie des résumés exhaustifs d'ordre I intervient.

Si en outre les échantillons sont eux-mêmes très grands (cas très courant en pratique) les lois-limites de Laplace-Gauss et la considération de l'estimation du maximum de vraisemblance entrent en jeu.

Le problème de premier plan des sondages est l'estimation de la moyenne arithmétique $\bar{x}$ de la population, par celle $\bar{X}$ de l'échantillon.

Si on a par ailleurs des raisons de penser que la distribution est, disons de Galton-Mac Alister-Gibrat, on pourra employer au lieu de $\bar{X}$ un estimateur X', tel que $\mathcal{V}X' < \mathcal{V}\bar{X}$ en particulier l'estimateur du maximum de vraisemblance X°, tel que

$$\mathcal{V}\ X^\circ = 1/nH$$

C'est là (on l'a dit) un cas exceptionnel; il est de pratique courante de réduire la perte d'information par ce qu'on connaît sur la structure de la distribution, la loi restant non spécifiée.

Exemples - Sondage stratifié.

$$X = x_{hi} = \bar{x}_h + \vartheta_{hi} \qquad \text{avec} \qquad \mathcal{E}\,\vartheta_h = 0$$

sans spécifier les lois centrées ϑ_h et connaissant les effectifs de strate ν_h, on abaisse la variance dans le rapport :

$$\Sigma\ \nu_h\,\sigma_h^2 / \Sigma\ \nu_h\,\sigma_h^2 + \Sigma\,\nu_h\,(\bar{x}_h - \bar{x})^2$$

dans le cas où les fractions sondées sont les mêmes dans chaque strate.

Si en outre les écarts-types de strate σ_h sont connus, on pose :

$$X = \bar{x}_h + \sigma_h \vartheta_h \qquad \text{avec} \qquad \mathcal{E}\vartheta_h = 0, \quad \mathcal{V}\vartheta_h = 1;$$

sans spécifier autrement les ϑ_h, le sondage "à la Neyman" abaisse la variance dans le rapport :

$$(\Sigma \nu_h \sigma_h)^2 / \nu \Sigma \nu_h \sigma_h^2$$

Autre exemple : Estimation par une droite de régression.

En postulant que X a une structure telle que :

$$X = Y + \vartheta \qquad \text{ou} \qquad X = Y + \sigma(y)\vartheta \qquad \text{avec} \qquad \mathcal{E}\vartheta = 0$$

on sait abaisser la variance dans le rapport $(1 - \rho^2)$ (ρ étant le coefficient de corrélation entre X et Y, et $\bar{y}$ étant supposé connu.

Dans chacun de ces cas, dès qu'on spécifie que les variables ϑ sont laplace-gaussiennes, ou simplement suivent des lois admettant la moyenne arithmétique comme résumé exhaustif, il devient impossible d'abaisser davantage la variance, du moins à effectif constant.

Bref il est utile de se souvenir qu'il existe une loi de distribution alors même qu'on en ignore tout.

B - De même il est permis de chercher à tirer parti d'une théorie de l'information qui suppose connues :

- la loi de distribution de X (à quelques ζ près);
- la loi de probabilité a priori des ζ.

Le tout est de savoir quel profit tirer du point de vue suivant lequel ζ cesse d'être une constante inconnue pour devenir une variable aléatoire.

Lorsqu'on cherche à appliquer le calcul des probabilités à l'analyse de données naturelles, il est bien normal qu'on ait quelque hésitation sur le choix du schéma d'urne à employer pour introduire le hasard dans le raisonnement. En revanche, lorsqu'on opère sur des données provenant de tirages au sort dont on a monté soi-même tous les mécanismes (et c'est le cas des méthodes modernes de sondage), on est en droit de douter de l'utilité et de la légitimité de l'introduction de variables aléatoires supplémentaires.

Au reste on a repris l'exemple du § X (du à Lindley) pour chercher à en tirer parti. Considérons une distribution laplace-gaussienne (μ, σ^2) de moyenne μ et d'écart-type σ, dont on extrait un échantillon

bernoullien d'effectif n; sa moyenne $\overline{X}$ suit a priori la loi laplace-gaussienne $(\mu, \sigma^2/n)$. Si μ est inconnu (mais σ connu) et si l'on connaît $\overline{X}$ (c'est-à-dire un certain échantillon), il est assez naturel de considérer la variable aléatoire auxiliaire M de distribution $(\overline{X}, \sigma^2/n)$ comme représentant l'estimation du paramètre inconnu μ. Cette dernière distribution est une probabilité a posteriori (connaissant l'échantillon). La loi de Bayes donne inversement une loi de probabilité a priori de M (et la loi de probabilité a posteriori de l'échantillon pour toute valeur donnée de M).

Pour alléger le calcul, on s'assurera de l'identité de deux lois gaussiennes à un nombre quelconque de variables en identifiant :

- d'une part les exposants de l'exponentielle (au facteur $-\frac{1}{2}$ près);

- d'autre part l'autre facteur de l'élément différentiel (à un multiple près de $1/\sqrt{2\pi}$).

La loi de Bayes s'écrit dans le cas d'un seul tirage :

$$p(x_1) \, . \, p(M \mid x_1) = p(M) \, . \, p(x_1 \mid M)$$

Identification : $\frac{(x_1 - \mu)^2}{\sigma^2} + \frac{(x_1 - M)^2}{\sigma^2} = \frac{(M - \mu)^2}{2\,\sigma^2} + \frac{2}{\sigma^2}\left(x_1 - \frac{M + \mu}{2}\right)^2$
ce qui fait bien apparaître la loi a priori $(\mu, 2\sigma^2)$ pour M.

Dans le cas de deux tirages, on a :

$$p(M) . p(x_1, x_2 \mid M) = p(x_1, x_2) \, . \, p(M \mid x_1, x_2)$$

avec $p(x_1, x_2) = p(x_1) \, . \, p(x_2) = p(x_1 - x_2) \, . \, p(x_1 + x_2)$

correspond aux égalités suivantes d'exposants :

$$(x_1-\mu)^2 + (x_2-\mu)^2 + \left(M - \frac{x_1 + x_2}{2}\right)^2 . \ 2 = \frac{3}{2}(x_1^2+x_2^2) + x_1x_2 + 2(M^2+ \mu^2)$$

$$- 2(x_1+x_2)(M+\mu) = \left[(x_1+x_2 - M - \mu)^2 + \frac{1}{2}(x_1 - x_2)^2\right] + (M - \mu)^2$$

où le crochet correspond à $p(x_1, x_2 \mid M)$ et $(M - \mu)^2$ à p(M).

Dans le cas de n tirages, il vient :

$$S(x_i - \mu)^2 + n(M - \overline{X})^2 = \left[S\left(x_i - \frac{M+\mu}{2}\right)^2 + n\left(\overline{X} - \frac{M+\mu}{2}\right)^2\right] + (M - \mu)^2 \ \frac{n}{2}$$

avec la même décomposition en $p(x_1, x_2, \ldots x_n | M) \cdot p(M)$.

Conclusion.

La loi de Bayes conduit pour M à une loi de probabilités a priori (avant tirage) : laplace-gaussienne $\left(\mu\ ,\ \frac{2\sigma^2}{n}\right)$.

Il en résulte que cette loi a priori ne serait pas la même quand on tire 1, 2, ... n boules, ce qui est absurde. Autrement dit :

Il n'est pas possible de représenter l'estimation (séquentielle) de μ par une variable aléatoire gaussienne M, constamment centrée en $\bar{X}$ et de variance σ^2/n, lorsqu'on suppose les variables X de valeur centrale μ inconnue mais fixe (comme c'est le cas lorsqu'on sonde une urne); on en conclut qu'il n'y a pas lieu de conserver l'aléatoire M.

Ceci ne signifie pas que, dans aucun cas, la loi de Bayes ne puisse avoir d'application. Mais nous la réserverions à des cas où il n'y a pas eu tirage au sort effectif (échantillon formé des premiers arrivants, par exemple).

XII - SUR L'INFORMATION SELON SCHÜTZENBERGER -

Schützenberger a établi (dans sa thèse, Pub. Inst. Stat. Uni. Paris, III, 1-2, page 43) le théorème suivant :

"Toute information(1) est la valeur moyenne, étendue à l'ensemble des états, de la résultante de l'application d'un opérateur linéaire S sur le logarithme de la probabilité a priori de chaque état".

En particulier avec S = - 1, on a l'information de Shannon;

avec $S = -\frac{\partial^2}{\partial \zeta^2}$, on retrouve l'information de Fisher,

etc.

D'où la question : à quel opérateur linéaire S correspondrait la "perte d'information" et en particulier la variance ?

1/ Tout d'abord les "états" évoqués dans une théorie volontairement très générale peuvent être compris de plusieurs manières.

----

(1) On notera qu'il existe de nombreux cas de pseudo-information, c'est-à-dire d'expressions qui, sans être des informations, sont de la forme :

$$p_i \ S \log p_i$$

Limitons-nous d'abord au sondage exhaustif dans une urne unique. A une taille n donnée de l'échantillon correspond :

- un plan de sondage π;

- C_{ν}^{n} = m échantillons distincts, donc m valeurs d'un estimateur Z;

- m^2 couples d'échantillons distincts, donc m^2 valeurs $\frac{1}{2}(Z_i - Z_j)^2$ ou plus généralement (ij), avec

$$(ii) = 0; \qquad (ij) = (ji)$$

etc.

a) Un "état" peut donc signifier un échantillon, un couple d'échantillons, etc., l'information (perdue) étant celle du plan de sondage tout entier.

En pareil cas, les probabilités p_i sont égales entre elles, - à 1/m pour la première interprétation, à $1/m^2$ pour la seconde, etc.

Si par "état" on entend "couple", peut-on considérer $\frac{1}{2}(Z_i - Z_j)^2$, - plus généralement (ij), - comme un S $\log(p_{ij})$.?

b) Mais par "état" on peut entendre encore une chaine d'échantillons de taille $(\nu - 1)$, $(\nu - 2)$, ... $(n + 2)$, $(n + 1)$, n, telle qu'on perde à chaque étape une unité de sondage de l'échantillon. Si les probabilités sont encore égales entre elles, leur expression est devenue :

$$n!/\nu!$$

Au lieu d'une seule chaine, on peut d'ailleurs en considérer deux simultanément. Mais, de toute manière nos "états" sont équiprobables.

L'information sera obtenue en divisant la somme des (S $\log p_i$) par le nombre d'états.

2/ Que faut-il comprendre alors par résultante de l'application d'un opérateur linéaire S ?

Considérons deux plans de sondage de tailles n et n'; soit π et π'.

Supposons que, n étant plus grand que n', on passe d'un "état" de π à un état de π', donc d'une probabilité p à une probabilité p' = pq.

Par exemple :

$$p = n!/\nu! , \qquad p' = n'!/\nu! , \qquad q = n'!/n!$$

d'où :

$$\log p = \log n! - \log \nu! , \qquad \log p' = \log n'! - \log \nu! ,$$

$$\log q = \log n'! - \log n!$$

a) On voit alors que $S \log p_i$ ne peut désigner $(Z - \zeta)^2$, pour la raison qu'on n'a pas :

$$(Z' - Z)^2 + (Z - \zeta)^2 = (Z' - \zeta)^2$$

l'égalité des deux membres supposant $(Z' - Z).(Z - \zeta) = 0$.

Il en résulte que "état" ne peut signifier "échantillon" avec lequel on calculerait Z (ou Z').

b) Désignons au contraire par état un couple (E_i, E_j) d'échantillons, chacun de taille n, et passons de là (en perdant n' - n unités pour chacun) à un couple (E'_i , E'_j) chacun de taille n'.

On voit cette fois qu'on peut poser :

$$2S = (Z'_i - Z'_j)^2 - (Z_i - Z_j)^2$$

expression dont le second terme se réduit à 0 pour $Z = \zeta$ (c'est-à-dire $\pi = U$).

En résumé : Par état on peut comprendre le couple d'échantillons; et par résultante de l'application de l'opérateur linéaire S sur log p, on peut entendre :

$$[(Z'_i - Z'_j)^2 - (Z_i - Z_j)^2] /2$$

ou (plus généralement) toute fonction G(I', J') - G(I, J) telle que G(J, I) = G(I, J), G(I, I) = 0.

Généralisation.

Pour les besoins par exemple de la mesure de la perte d'information sur l'estimation d'une perte d'information, on pourra considérer comme état un groupe de 4 échantillons et pour

$$S \log p \doteq (Z'_i - Z'_j)^2 (Z'_h - Z'_k)^2 - (Z_i - Z_j)^2 (Z_h - Z_k)^2$$

ainsi que des expressions plus générales.

Cas de plans de sondage plus complexes.

La transposition est immédiate. En particulier les tirages avec probabilités inégales (associés à des estimateurs sans biais) modifient le calcul de l'espérance mathématique de S log p, mais non S log p.

XIII - AUTRE INTERPRETATION DE LA PERTE D'INFORMATION (T. CR 4 1958) -

Revenons sur le théorème de Schützenberger évoqué au § XII.

Soit i, j, deux échantillons au hasard (distincts ou non) du même plan de sondage et F une fonction positive, telle que :

$$F(i, j) = F(j, i), \qquad F(ii) = 0$$

La probabilité d'obtenir le couple ij est $1/m^2$ s'il existe m échantillons équiprobables pour ce plan de sondage.

Appelons écart D(ij) entre les deux échantillons ij l'expression :

$$D(ij) = \left[\exp\,[-F(ij)]\;\; m^2 \right]^{-1}$$

et densité le rapport $m^{-2}/D(ij) = p(ij)$

$$\text{soit} \qquad q(ij) = 1 - \exp\,[-F(ij)]$$

C'est bien une probabilité (comprise entre 0 et 1); posons de même :

$$p(ij) = \frac{d}{dF}\, q(ij) = \exp\,[-F(ij)]$$

d'où

$$-\log p(ij) = F(ij).$$

Appelons enfin $\mathcal{E}\,[-\log p(ij)] = \mathcal{E}\,[F(ij)] = \mathcal{E}(pi)$ la perte d'information. On retrouve à la fois la définition de Shannon et la nôtre.

Donc, tout se passerait comme si la densité de probabilité de l'écart entre les deux échantillons ij était de la forme :

$$A \,.\, \exp\,[-F(ij)\,.\,B]$$

où A et B désignent deux constantes (positives).

En particulier : Cas de la variance.

Pour estimer sans biais le paramètre ζ, on dispose de

$$Z(i) \quad \text{et} \quad Z(j)$$

c'est-à-dire deux estimations, sur échantillons tirés au sort (i) et (j), avec l'estimateur sans biais Z.

Soit $$F(ij) = [Z(i) - Z(j)]^2/2$$

On sait que : $$(pi) = \mathcal{V} Z = \mathcal{E} F$$

Pour retrouver l'information de Shannon, il faut donc que Z(i) - Z(j) ait une densité de probabilité de Laplace-Gauss.

Lorsque l'échantillon est grand, les moyennes sur échantillon admettent des lois-limite de Laplace-Gauss; de même les estimations du maximum de vraisemblance; etc. Supposons que ce soit le cas pour Z. La différence Z(i) - Z(j) suit elle-même une loi de cette nature.

Ainsi, lorsque l'estimateur admet une loi-limite de Laplace-Gauss, il y a coïncidence entre l'information de Shannon et la perte d'information dans le cas des grands échantillons.

Mais si l'on écarte cette interprétation asymptotique, on est réduit à supposer que le plan de sondage est bernoullien et la distribution de base laplace-gaussienne, ce qui ne présente plus beaucoup d'intérêt pratique (§ X ci-dessus).

XIV - SUR L'INFORMATION SELON FISHER -

Finalement l'opérateur linéaire S de Schützenberger qui est utilisé en sondage, n'a donc rien de commun avec celui ($\partial^2/\partial\zeta^2$) de Fisher.

La relation simple entre σ^2/n et n/σ^2 (avec une loi de Laplace-Gauss) apparaît comme accidentelle. Essayons de l'expliquer.

Limitons-nous au cas de l'estimation d'un seul paramètre ζ, par Z.

Soit F la loi de probabilité de Z, et G celle des variables X_1, X_2, ... X_n indépendantes au début et qu'on lie par une valeur donnée à Z. On désigne par f_i la loi de probabilité de X_i. Il vient :

$$\sum_{(n)} \log f_i = \log F + \log G$$

En dérivant deux fois par rapport à ζ, en prenant les espérances mathématiques des deux membres et en changeant les signes, il vient :

$$\text{Information de Fisher} = \mathcal{E}\left[-\frac{\partial^2}{\partial\zeta^2}(\log F)\right] + \mathcal{E}\left[-\frac{\partial^2}{\partial\zeta^2}(\log G)\right]$$

$$= \mathcal{E}\left[\frac{\partial}{\partial\zeta}\quad(\log F)\right]^2 + \mathcal{E}\left[\frac{\partial}{\partial\zeta}\quad(\log G)\right]^2$$

Lorsque la loi f(X) admet un résumé exhaustif d'ordre 1 par rapport à ζ, le deuxième terme du deuxième membre disparaît; il subsiste autrement.

Lorsque n est grand, $\overline{X}$ admet une loi-limite de Laplace-Gauss, et on va s'en tenir aux estimateurs Z possédant cette même propriété.

Par conséquent on aura :

$$\mathcal{E}\left[-\frac{\partial^2}{\partial\zeta^2}\log F\right] = \frac{n}{\sigma^2} + \varepsilon$$

où σ n'est pas l'écart-type des X mais celui des Z pour n = 1 (ou à défaut pour n = r).

Considérons alors un couple d'échantillons de taille n, soit (1), (2).

Si Z(1) et Z(2) sont sans biais, la variable Z(1) - Z(2) est indépendante de ζ. De plus, comme n est grand, elle est ℓ laplace-Gaussienne, comme la variable Z(1) + Z(2) dont elle est indépendante.

Le moment du 2ème ordre de Z(1) - Z(2), en particulier, ne peut dépendre de ζ. Au contraire, l'information de Fisher est une fonction de ζ, constante dans le cas particulier où log f est un polynome du 2ème degré en ζ (exclusivement).

Pour le couple (1) (2), l'information de Fisher est :

$$\mathcal{E}\left[-\frac{\partial^2}{\partial\zeta^2}\log F(1+2)\right] + \mathcal{E}\left[-\frac{\partial^2}{\partial\zeta^2}\log G(1) - \frac{\partial^2}{\partial\zeta^2}\log G(2)\right]$$

où le premier terme est voisin de $2n/\sigma^2$. On sait que σ^2 ne dépend pas de ζ; c'est donc dans le second terme qu'est localisée l'influence de ζ.

Finalement c'est cette présence de

$$\mathcal{E}(-\partial^2 \log G/\partial\zeta^2) = \mathcal{E}(\partial \log G/\partial\zeta)^2$$

à côté de $\mathcal{E}(\partial \log F/\partial\zeta)^2$ qui rend l'information de Fisher (pour n grand) au moins égale à l'inverse de $\mathcal{V}(Z)$.

XV - DERNIERES REMARQUES -

Les informations de Lindley ou de Fisher sont bien entendu des solutions de l'équation de Récurrence de Schützenberger. Mais l'espérance mathématique qui y figure a, pour l'Information de Lindley, une signification fort différente de la nôtre; alors que nous lui donnons le même sens que Fisher. Le raisonnement que nous avons fait, à la IIIème partie du Chapitre IV n'est valable, il est vrai, que pour des pertes d'information finies; mais il serait possible de l'adapter au cas présent. Ainsi est-on conduit à poser :

$$\vartheta(\pi') - \vartheta(\pi) = n$$

$$1/H = E\, g(i, j) = \gamma$$

avec deux sondages bernoulliens (π') et (π), d'effectifs :

$$\lambda \text{ et } \lambda + n$$

avec
$$2g(i, j) = \left[\frac{\partial}{\partial \zeta} \log \frac{f(x_i, \zeta)}{f(x_j, \zeta)}\right]^{-2}$$

Ainsi s'achève cet exposé qui, à notre regret, malgré sa longueur, laisse encore bien des points dans l'ombre.

BIBLIOGRAPHIE

Reference

A.1. ANSCOMBE - Sequential Estimation. J. Royal Stat. Society - Series B. 1953. 1.

BAR G. A. BARNARD - The theory of information - J. Royal Stat. Society - Series B. 1951, p. 46.

BA 1. D. BASU - On symmetric estimators in point estimation with convex weight functions.-SANKHYA 12, p.45 (décembre 1952).

BA 2. D. BASU - On the optimum character of some estimators used in multistage sampling problems - SANKHYA, 13, p. 3 63 (juin 1954).

CHA 1. CHARTIER F. - Note sur les fluctuations aléatoires d'une variance estimée - Bulletin d'information de l'INSEE - n°3 mars 1957 - p. 31-40.

CHA 2. CHARTIER F. - Sur l'influence d'une stratification aléatoire lors de l'estimation du total d'une population - Journal de la Société de Statistique de Paris 97, 4-5-6. 1956, p. 121-129 .

CR.1. Harald CRAMER - Mathematical Methods of Statistics - Princeton, 1946.

CR.2. Harald CRAMER - Sur un nouveau théorème-limite de la théorie des probabilités - 1ère partie du fascicule : Les sommes et les fonctions de variables aléatoires - Actualités Scientifiques n°736; Hermann 1938.

DA.1. G. DARMOIS - Sur les lois de probabilités à estimation exhaustive CR. 200 1935 - p. 1265.

DA.2. G. DARMOIS - Méthodes d'estimation - Actualités scientifiques Hermann 1936.

DA.3. G. DARMOIS - Sur l'estimation des grandeurs par leurs mesures (Annuaire du Bureau des Longitudes. 1952).

Référence

DU. D. DUGUE - Application des propriétés de la limite au sens du calcul des probabilités à l'étude de diverses questions d'estimation. - Journal de l'Ec. Polytechnique 1937 - p. 305.

FE FERON - Information, Régression, Corrélation - Thèse pour le doctorat ès Sciences, Paris 1954.

FIN FINNEY - On the distribution of a variate whose logarithm is normally distributed, page 155. Suppl^{t} of the J. of the Royal Stat. Society VII n°2, 1941.

FIS 2 FISHER (Sir Ronald) - On the mathematical foundations of theoretical Statistics - Phil. Trans. Roy. Soc. London - Series A. 222 p. 309 - (1922).

FIS 1 FISHER (Sir Ronald) - Theory of statistical estimation - Proceedings Cambridge Phil. Society 22 (1925) p. 700.

FIS 3 FISHER (Sir Ronald) - Statistical methods and Scientific Inference - Oliver & Boyd. - 1956 - p. 148, etc.

FOR R. FORTET - Calcul des probabilités - CNRS 1950.

GE GEISSER (Seymour) - A note on the normal distribution. - The Annals of Math. Statistics - septembre 1956 - p. 858 -
(Geary. J.R.Stat.Soc.Suppl. Vol.3 (1936) n°2.
(Réf.:)Lukacs. Ann.Math.Stat. Vol. 13 (1942) p. 91.
(Daly. Ann. Math. Stat. Vol. 17 (1946) p. 71).

KH Abbas. G. KHADJENOURI - Contribution à la théorie mathématique de l'échantillonnage - Thèse pour le Doctorat ès Sciences - PARIS 1956 -

LI 1. D.V. LINDLEY - Statistical Inference. - J. of the Royal Stat. Society - 1953 - Séries B. XV n°1 - p. 30.

LI 2. D.V. LINDLEY - On a measure of the Information provided by an experiment - The Annals of Math. Statistics - 27 n°4 déc. 1956 - p. 986.

LI.3. D.V. LINDLEY - Binomial sampling schemas and the concept of information - Biometrika Vol. 44 n°1-2 - p. 1957.

N.1. Jerzy NEYMAN - On the two different Aspects of the Representative method : the method of Stratified Sampling and the method of Purposive Sampling. - Journal of the Royal Statistical Society 1934 - p. 558-625.

N.2. Jerzy NEYMAN - Contribution to the Theory of Sampling Human Population - J. Amer. Stat. Assoc. 33 (1938) p. 101-116.

N.3 Jerzy NEYMAN - Sur une famille de tests asymptotiques des

Référence

hypothèses statistiques composées - Madrid - Revista Trabajos de Estadistica - V. II. 1954.

SC. M. P. SCHÜTZENBERGER - Contribution aux applications statistiques de la théorie de l'information (Thèse pour le doctorat ès-Sciences-Paris 1954) Publications de l'Inst. de Stat. de l'Université de Paris - III. 1. 2. 1954.

ST. STUART (Alan) - The measurement of estimation and test efficiency - Session 1957 de l'Inst. Inst. Stat. - papier n°62.

TU. TUKEY - Variances of Variance component - Ann. Math. St. 1956, p. 722.

UT. J.E.G. UTTING & Dorothy COLE - Sample Surveys for the Social Accounts of the Household Sector - Bulletin of the Oxford University Institute of Statistics - janv. 1953.

WAL. Abraham WALD - Statistical Decision Functions - New York 1950.

Z 1. ŽARKOVIĆ - Sampling Control of Literacy Data. Journal of the Am. Stat. Assoc. 1954 - pp. 510-519.

T.CR Pierre THIONET - Notes aux comptes-rendus Acad. Sci. :

1. - Sur la variance de l'estimation d'une variance (Vol. 245, 1957, p. 2168).
2. - Une généralisation de la variance d'échantillonnage dans le cas de tirages exhaustifs d'une urne (Vol. 245, 1957, p. 2464).
3. - Représentation topologique des sondages (Vol. 246, 1958 p. 46).
4. - Sur les rapports entre divers concepts d'information (Vol. 246, 1958, p. 123).
5. - Sur les pertes d'information qui sont des fonctions de risque (Vol. 246, 1958, p. 128).
6. - Sur les pertes d'information imputables à certaines estimations biaisées (Vol. 246, 1958, p. 536).
7. - Sur une théorie générale des pertes d'information par sondage (Vol. 246, 1958, p. 692).

T.IIS Pierre THIONET - Erreurs d'échantillonnage - fonctions de risque et information perdue - Session 1957 de l'Inst. Int. de Stat. - Papier n°64 - Bulletin d'Inf. de l'INSEE, sept-nov. 1957 - p. 51-64.

T.1. Pierre THIONET - Méthodes statistiques modernes des administrations fédérales aux Etats-Unis. - Actualités Scientifiques. Hermann 1946.

Référence

T.2. Pierre THIONET - L'école moderne de statisticiens italiens - Journal de la Société de Statistique de Paris (1945 n°11-12 1946 n°1-2).

T.3. Pierre THIONET - La théorie des sondages - Etude théorique n°5 de l'INSEE - Paris (I. N) 1953.

T.4. Pierre THIONET - Application des méthodes de sondage aux enquêtes statistiques - Etude théorique n°6 de l'INSEE - Paris (I. N.) 1953).

T.5. Pierre THIONET - Théorie des sondages - Cours de l'Ecole d'application de l'INSEE (année scolaire 1955-56).

T.6. Pierre THIONET - En s'exerçant à la théorie des sondages - Bulletin d'information de l'INSEE - sept-octobre 1954.

T.7. Pierre THIONET - Recherche d'une fonction de coût applicable aux enquêtes par sondage de l'INSEE. - Revue de l'Institut International de Statistique 1954 - p. 33.

T.8. Pierre THIONET - Décisions à propos de sondages - Revue de Statistique appliquée - 1955 p. 71-80.

T.9. Pierre THIONET - Le problème théorique du plan d'échantillonnage - Journal de la Société de Statistique de Paris - 1948 - mars-avril.

T.10. Pierre THIONET - a) Considérations théoriques à propos de sondage (1955).
b) Notion d'Information et théorie des sondages (1956).
(Documents intérieurs au S. E. E. F.).

T.11. Pierre THIONET - Théorie des décisions statistiques et théorie des sondages.
(Bull. d'Inf. de l'INSEE - sept-oct. 1956, p. 1. 6).

TABLE DES MATIÈRES

LA MÉTHODE STATISTIQUE EN MÉDECINE : LES ENQUÊTES ÉTIOLOGIQUES

Daniel SCHWARTZ

Exposé fait au Séminaire de Statistique le 10 Novembre 1959

On se propose de rechercher dans quelle mesure un facteur x intervient causalement dans le déterminisme d'une maladie m au sein d'une population humaine : par exemple l'usage du tabac dans le cas du cancer broncho-pulmonaire.

Ce problème peut théoriquement être abordé :

- soit par la voie expérimentale : examen de 2 groupes comparables, obtenus par tirage au sort, dont l'un sera soumis au facteur x et l'autre non. Cette façon de faire est le plus souvent inapplicable pour des raisons matérielles ou morales; elle ne répond d'ailleurs pas exactement à la question posée : avec ce procédé autoritaire le cancer du poumon pourrait bien frapper, dans le groupe fumeur, des sujets particulièrement vulnérables, qui dans les conditions spontanées ne fumeraient pas;

- soit par la voie de l'observation : on cherche s'il existe, dans la population générale, une association entre l'exposition au facteur x et l'apparition de la maladie m. Cependant l'association ne permet pas de conclure à la causalité, du fait que l'exposition au facteur x est aléatoire, et liée à de nombreux facteurs, parmi lesquels peut se trouver la vraie cause : si l'usage du tabac résulte d'un psychisme déterminé, la consommation de tabac élevée des cancéreux ne serait-elle pas seulement l'indice d'un psychisme particulier, qui serait la cause de ce cancer ? La voie de l'observation ne saurait donc en théorie rien apporter au problème étiologique. En fait elle a permis, dans certains cas qui seront développés en fin de cet exposé, d'aboutir à une forte présomption causale.

Cependant même si, renonçant à l'interprétation causale, on se contente plus modestement d'étudier la relation d'association, il faut préciser d'emblée que cette association à l'état brut est souvent inintéressante : ainsi, dans l'exemple qui vient d'être cité, on doit s'attendre à observer beaucoup plus souvent le cancer broncho-pulmonaire chez les fumeurs que chez les non fumeurs, par le seul fait que l'âge

----

*Unité de Recherches Statistiques de l'Institut National d'Hygiène (Institut Gustave Roussy).

moyen est beaucoup plus élevé dans le premier groupe que dans le second, qui comprend jusqu'aux nouveaux-nés; il n'est intéressant de comparer les fumeurs et les non fumeurs qu'à âge égal; on arrive ainsi à la notion d'une association corrigée de l'influence de l'âge. Ce problème pourra être abordé, soit en examinant seulement une population d'âge donné (étude en population homogène), soit par une étude en population hétérogène, couvrant un assez large intervalle d'âges, l'influence de l'âge étant éliminée par un procédé statistique.

Les mêmes considérations s'appliquent naturellement à des facteurs autres que l'âge : sexe, peut-être milieu d'habitation, niveau social... d'une manière générale à tous les facteurs liés à la consommation de tabac. La liste de ces facteurs peut être longue et inconnue. En fait on peut assez raisonnablement distinguer 2 étapes :

- dans une première étape, on étudie l'association, d'une part à l'état brut, mais simultanément en la corrigeant de l'influence du sexe et de l'âge, et éventuellement d'un nombre extrêmement réduit de facteurs liés fondamentalement à l'exposition au facteur x.

- dans une deuxième étape, on corrige l'association pour tous les autres facteurs liés à l'exposition au facteur x. Cette étape peut être menée plus ou moins loin. Elle est sans fin ...

On peut admettre, sans trop d'arbitraire, que la limite entre ces 2 étapes définit le moment où se termine l'étude du rôle étiologique du facteur, et où commence la recherche d'une interprétation causale.

L'une et l'autre peuvent être abordées : soit par l'examen d'une population homogène, où tous les facteurs en cause sont constants, soit par l'examen d'une population hétérogène, où le rôle de ces facteurs est éliminé par un procédé statistique.

Ces considérations définissent le plan de notre exposé.

I - PRINCIPE DE L'ENQUETE ET MODE D'ECHANTILLONNAGE -

a) UN PREMIER EXEMPLE (enquête prospective).

Dans cet exemple, on constitue un échantillon représentatif de la population générale; on note si les sujets sont exposés ou non au facteur x, et on enregistre par la suite tous les cas de la maladie m qui se produisent.

De telles enquêtes ont été conduites notamment pour étudier le rôle de l'obésité et de l'hypertension dans l'étiologie de la maladie coronarienne (18), de la rubéole des femmes enceintes dans la produc-

tion de malformations chez l'enfant (36), de la consommation de tabac dans le déterminisme de diverses maladies, en particulier le cancer des voies aéro-digestives supérieures (24, 25, 32).

L'échantillon examiné doit être représentatif. Cependant cette exigence n'est pas toujours réalisable, et dans certains cas on a choisi un groupe plus facile à suivre, par exemple la totalité du corps médical (24) ou un groupe de sujets pensionnés de l'Etat (25), faisant l'hypothèse d'une stabilité de l'association étudiée (si le tabac est dangereux pour le corps médical, il l'est vraisemblablement d'une manière générale). En tout état de cause, si on ne cherche pas à extrapoler, on dispose au moins de résultats valables pour une population bien définie.

L'échantillon peut être représentatif d'une population homogène : l'enquête (32) sur les fumeurs portait sur les sujets de sexe masculin, de race blanche, d'âge compris entre 50 et 69ans. Par contre, l'enquête (24) sur le corps médical portait sur une population hétérogène en âges, et l'étude d'association à âge donné nécessita une correction par un des procédés qui seront exposés plus loin.

L'exposition au facteur x est connue par un examen ou un questionnaire, nécessairement très réduit en raison du grand nombre des sujets; quant à la maladie, elle est enregistrée lorsque c'est possible; cependant on doit souvent se contenter de l'information du décès, substituant à l'étude d'une maladie l'étude de la mort par cette maladie.

Une enquête "prospective" de ce genre présente d'indiscutables avantages : elle fournit, comme on le verra plus loin, une information complète sur le rôle du facteur; elle évite le recours à des groupes témoins critiquables; l'interrogatoire a lieu à un moment où le sort du sujet n'est pas connu, ce qui en garantit l'impartialité. Enfin il est possible d'étudier simultanément les diverses maladies imputables au facteur.

Mais le nombre de sujets exigé est considérable, dès que la fréquence de la maladie est faible : on a suivi près de 200 000 sujets pendant plusieurs années dans les enquêtes (25) et (32). De telles enquêtes sont donc rarement réalisables.

b) UN DEUXIEME EXEMPLE (enquête rétrospective).

Dans ce deuxième exemple, on constitue un échantillon de sujets atteints de la maladie, et un échantillon témoin qui en est indemne, et on les compare pour la proportion de sujets exposés au facteur x. L'échantillon de malades doit être représentatif; si ceci a pu être réalisé dans de rares cas, notamment dans une enquête sur les cancers et

leucémies de l'enfant qui a pu englober tous les cas (du moins tous les cas mortels) pendant une période donnée (62), on se contente en général d'un mode de recrutement commode, par exemple des cas rencontrés à l'hôpital et dans certaines villes, admettant ici encore la stabilité de l'association.

C'est alors la définition du groupe témoin qui devient difficile : il doit être obtenu par tirage au sort parmi les sujets indemnes de la maladie dans "la population d'où provient l'échantillon de malades". Si on choisit pour le groupe malade les cas hospitaliers, on admettra que cette population est "la clientèle hospitalière", c'est-à-dire une catégorie de sujets que leur condition (sociale, familiale, psychologique ...) rend candidats à l'hôpital en cas de maladie.

En réalité il n'existe pas une clientèle hospitalière en général, mais une clientèle par maladie : plus celle-ci est grave, plus l'hôpital recrute à une grande distance et dans des classes sociales de niveau élevé; c'est donc la clientèle spécifique de la maladie m qu'il faut échantillonner pour constituer le groupe témoin. Pratiquement on forme le groupe témoin avec les cas hospitaliers d'une maladie m' de gravité comparable à m; ce procédé n'est acceptable que si la maladie m' a frappé cette clientèle "au hasard", c'est-à-dire si elle ne présente pas de relation avec le facteur x (ce qui va de soi : il ne faut évidemment pas que la maladie m' soit liée à une sous-exposition ou à une sur-exposition au facteur). On choisira par exemple, pour l'étude d'un cancer, des témoins atteints d'autres cancers, ou d'autres maladies graves; ou bien, partant du groupe de sujets venus consulter pour une tumeur, dont ils ne savent si elle est bénigne ou maligne, on les divisera après coup en cancers (maladie à étudier) et tumeurs bénignes (témoins), étant à peu près assuré que les mêmes facteurs d'échantillonnage ont dirigé les uns et les autres vers l'hôpital.

La difficulté de trouver un groupe témoin correct conduit en général à choisir plusieurs groupes témoins, qu'on justifie par une comparaison mutuelle : par exemple, dans l'enquête (60) sur le cancer broncho-pulmonaire, les 3 groupes témoins choisis (cancers autres que ceux des voies aéro-digestives supérieures, malades des services de médecine générale, accidentés) ont présenté le même niveau de consommation de cigarettes (alors que celui-ci était beaucoup plus élevé pour les cancers du poumon).

De toute manière, si on suppose que le groupe témoin ne représente pas correctement la population d'où provient le groupe malade, il devient nécessaire de corriger ce biais, soit par l'examen de populations homogènes (comparaison des malades et des témoins dans des groupes de milieu d'habitation, niveau social ... donnés), soit en

appliquant à l'échantillon de population hétérogène les corrections voulues, qui seront exposées plus loin.

Cette partie de l'analyse statistique doit être menée avec soin, si on veut éviter d'appeler association ce qui ne serait en fait que le résultat d'une inégalité d'échantillonnage entre les 2 groupes.

C'est pourquoi on tâche en général de corriger cette inégalité par un appariement; ce procédé, qui est un des avantages possibles de l'enquête rétrospective, consiste à chercher, pour chaque malade interrogé, un témoin comparable eu égard à certaines caractéristiques : celles-ci peuvent être, soit des facteurs d'échantillonnage (milieu d'habitation, niveau social...) soit des facteurs essentiels cités plus haut : sexe, âge ... Ainsi, dans l'enquête (60) sur l'étiologie du cancer broncho-pulmonaire, à chaque malade correspondait un témoin d'âge voisin (même tranche d'âge de 5 ans), interrogé à la même époque et si possible dans le même hôpital; dans l'enquête sur les cancers de l'enfant (62), les témoins étaient tirés au sort sur les registres de l'état-civil de la commune où était né l'enfant cancéreux, parmi les enfants de même sexe nés à la même date, ce qui assurait l'appariement par sexe, âge et lieu d'habitation.

Il va de soi que l'appariement doit être limité aux seuls facteurs dont le rôle est déjà connu; dès lors qu'on apparie en fonction d'un facteur, on annule pour ce facteur la différence entre les groupes malade et témoin, de sorte qu'on renonce à toute information sur son rôle dans l'étiologie.

L'exemple ainsi décrit d'enquête rétrospective comporte finalement bien des difficultés de principe; en contre-partie la conduite de l'enquête est infiniment plus aisée que dans les enquêtes prospectives, car il suffit de réunir un nombre relativement faible de cas. En outre, il devient alors possible de mettre en jeu un questionnaire détaillé, de sorte que ce n'est pas seulement le rôle d'un facteur qui est étudié, mais de plusieurs, voire de tous ceux dont l'influence étiologique est supposée.

Aussi ce genre d'enquête a-t-il tenté de nombreux chercheurs, qui l'ont utilisé pour des maladies aussi variées que la tuberculose (44), la maladie coronarienne (21, 28, 65), la cirrhose du foie (54), les malformations congénitales (45) et surtout le cancer. Dans ce seul domaine, on a ainsi étudié la relation entre la situation de famille et le cancer du sein (30, 37, 59, 63) ou du col de l'utérus (30, 34, 43, 68), l'étiologie du cancer de la vessie en fonction de l'usage du tabac (11, 19, 40) ou d'infections parasitaires (52), du cancer des voies aéro-digestives supérieures en relation notamment avec la consommation du

tabac et de l'alcool (38, 50, 55, 56, 58, 69, 71), des cancers gastro-intestinaux en relation avec l'usage des laxatifs (6), du cancer gastrique en relation avec les antécédents héréditaires (64) ou le groupe sanguin (5, 29) de la leucémie chez l'enfant en relation avec une irradiation de la mère pendant la grossesse (62). Rien que pour le cancer broncho-pulmonaire, on peut citer plus de 20 enquêtes rétrospectives (notamment 7, 22, 55, 60, 67 pour ne mentionner que celles qui portent sur au moins 500 cas de cancer et 500 témoins, et 31, 70 pour celles qui portent sur le sexe féminin).

c) CONSIDERATIONS GENERALES - CLASSIFICATION DES ENQUETES.

Les 2 modes d'enquête qui viennent d'être décrits sont très différents, et l'élément le plus apparent de cette différence est d'ordre chronologique : on s'attache à l'avenir des sujets dans un cas, au passé dans l'autre. Cependant le temps n'est ici qu'un caractère second, et il est bien plus judicieux de classer les enquêtes d'après le mode d'échantillonnage.

Nous adopterons à cet effet un modèle, représenté au tableau 2, où les sujets sont classés dans un tableau 2 × 2 en 4 catégories. Il s'agit là d'un modèle simplifié; en effet :

a) la raison pour laquelle les sujets non exposés peuvent contracter la maladie m n'est pas envisagée; cette raison peut être l'exposition à un facteur y au moins; il faudrait dans ce cas prévoir un modèle à au moins 3 dimensions (exposition au facteur x, exposition au facteur y, maladie m);

b) on pourrait étudier avec plus de précision le rôle du facteur x, en supposant plusieurs degrés d'exposition. Ce point sera parfois pris en considération dans les pages qui suivent;

c) il faut enfin préciser ce qu'on entend par "sujets atteints de la maladie m". Il peut s'agir de mortalité ou de morbidité, et, dans cette dernière éventualité, soit des nouveaux cas apparus pendant une période donnée, soit des cas existant à un moment donné (incidence et prevalence de la terminologie anglo-saxonne); ces aspects, - et d'autres qu'on peut imaginer - traduisent des moyens différents de mesurer la fréquence d'une maladie dans un groupe. Cette diversité est commune à bien des problèmes d'ordre statistique en médecine, et la source de bien des difficultés. Nous conserverons dans le tableau 2 la terminologie, à dessein vague, de "malades", en sachant qu'il y aurait lieu, pour chaque problème particulier, de formuler au départ une définition plus précise.

Si on adopte le modèle du tableau 2, les 2 variables, exposition au facteur x et atteinte par la maladie m, étant toutes deux aléatoires (puisqu'il s'agit uniquement d'observation, l'expérience étant exclue), c'est la nature de leur distribution - distribution contrôlée ou distribution aléatoire - qui permet de classer les types d'enquête. On obtient alors, avec White et Bailar (66), 3 types :

Type 1 - distribution aléatoire pour x et pour m : on constitue un échantillon représentatif de la population étudiée.

Type 2 - distribution contrôlée pour x, aléatoire pour m : on constitue 2 groupes représentatifs de sujets exposés et non exposés.

Type 3 - distribution contrôlée pour m, aléatoire pour x : on constitue 2 groupes représentatifs de sujets malades et non malades.

TYPE 1 (échantillon représentatif de la population étudiée).

C'est dans cette catégorie qu'entrent les enquêtes prospectives décrites plus haut. Toutefois le type 1 n'oblige aucunement à suivre les malades dans le futur, on peut très bien dans certains cas se référer au passé ou au présent des sujets. Naturellement, s'il s'agit du cancer du poumon, on ne saurait s'intéresser au passé, car les sujets ayant dans le passé développé cette maladie seront en majorité décédés, ce qui faussera l'échantillonnage; on ne peut pas davantage s'intéresser au présent, car le nombre de sujets atteints serait trop faible; force est donc de suivre les sujets dans le futur. Mais dans le cas d'une maladie non mortelle, et fréquente, rien n'empêche de considérer le passé ou le présent : on pourra par exemple étudier la relation entre l'éthylisme et les altérations artérielles du fond d'œil sur un échantillon de taille modeste, le facteur et le signe pathologique étant tous deux largement répandus.

Au type 1 se rattache la catégorie particulièrement intéressante des enquêtes de morbidité, qui indiquent les nouveaux cas de maladie apparus, pendant une période déterminée, dans la population entière d'une aire géographique déterminée, et constituent des enquêtes étiologiques possibles lorsque le facteur x est une caractéristique démographique connue par les statistiques de cette population : ainsi a-t-on pu étudier la relation entre les cancers génitaux de la femme et la situation de famille, dans 10 grandes villes des U.S.A. (26) et dans la totalité du Danemark (10).

TYPE 2 (un groupe de sujets exposés, un groupe de sujets non exposés).

Ce type peut permettre de suivre un nombre de sujets moins considérable que dans le type 1. Avec les notations des tableaux 1 et 2, la comparaison des 2 groupes fait intervenir une variance de forme

$\frac{m_0(1 - m_0)}{n_{0*}} + \frac{m_1(1 - m_1)}{n_{1*}}$; on peut, se fixant celle-ci, chercher les valeurs de n_{0*} et n_{1*} qui assurent l'effectif total ($n_{0*} + n_{1*}$) minimum. Si on suppose que la fréquence de la maladie ne sera pas beaucoup plus élevée dans le groupe exposé que dans le groupe non exposé ($m_1 \# m_0$), c'est en choisissant des effectifs égaux dans les 2 groupes qu'on obtient le minimum de sujets à suivre.

Dans le cas de l'enquête sur le rôle du tabac, un échantillonnage aléatoire de 200 000 sujets conduit à 30 000 non fumeurs et 170 000 fumeurs; il est certain qu'en constituant au départ 2 groupes d'effectifs plus équilibrés, on peut, pour une même précision, diminuer le nombre de sujets nécessaire; cela obligerait, par ailleurs, pour trouver davantage de non fumeurs, à organiser une prospection initiale plus étendue : peut-être serait-ce finalement plus compliqué, ceci dépend des difficultés relatives de la prospection initiale et de la surveillance ultérieure. La surveillance est en général difficile; la prospection initiale peut être aisée : s'il s'agit d'étudier la fréquence des cancers génitaux de la femme en fonction du nombre d'enfants, ce dernier renseignement sera facilement disponible, et on aura tout intérêt à constituer 2 groupes d'effectif équivalent de femmes avec ou sans enfants.

Le bénéfice du type 2 est d'autant plus considérable que l'exposition au facteur est plus rare (par exemple exercice d'une profession peu répandue). Si celle-ci se rencontre 1 fois sur 1 000, le coefficient de la variance serait, par millier de sujets, dans le type 1 (toujours dans l'hypothèse $m_0 \# m_1$), $\frac{1}{1} + \frac{1}{999} \# 1$; précision qui peut être obtenue dans le type 2 par $\frac{1}{2} + \frac{1}{2}$, donc avec 4 sujets; il suffit ainsi d'un nombre de sujets 250 fois plus faible.

A ces gains souvent très considérables, le type 2 permet d'ajouter encore un perfectionnement : lorsque l'exposition au facteur peut être divisée en plus de 2 classes hiérarchisées (0, 1, 2, ... enfants, ou non fumeurs, petits, moyens, grands fumeurs), si on suppose que l'effet du facteur croît en fonction de cette hiérarchie, on peut constituer 2 groupes, correspondant aux valeurs extrêmes (non fumeurs et très grands fumeurs, femmes sans enfants et mères de famille nombreuse). L'écart escompté entre les 2 groupes étant augmenté, on pourra se contenter d'effectifs plus faibles.

Enfin un cas extrême du type 2 est celui où on constitue seulement le groupe exposé au facteur, le groupe non exposé s'identifiant à la population générale : on a suivi, par exemple, un groupe d'ouvriers travaillant dans l'amiante, et comparé la fréquence observée de décès par cancer du poumon à celle de la population générale (23); on a de même

étudié la mortalité par cancer du poumon chez les sujets gazés, ou souffrant de bronchite chronique (9), la mortalité par cancer de l'estomac chez des personnes achlorhydriques ou atteintes d'anémie pernicieuse (3, 33, 51), la mortalité chez les radiologistes, pour les différentes causes de décès, et en particulier le cancer (17, etc.). Cette méthode suppose naturellement que les sujets exposés constituent, dans la population générale, un groupe suffisamment petit pour qu'on puisse confondre population non exposée et population générale. Par ailleurs, les comparaisons de mortalité ou de morbidité ne s'étendent évidemment qu'à sexe égal, âge égal, éventuellement milieu social égal, etc. ce qui exige les corrections d'usage.

TYPE 3 (un groupe de sujets malades, un groupe de sujets non malades).

C'est dans cette catégorie qu'entrent les enquêtes rétrospectives décrites plus haut. Elle permet de réduire les effectifs prévus par le type 1, tout comme le type 2, et pour des considérations symétriques, portant cette fois sur les effectifs des groupes malade et témoin. Le gain est obtenu en équilibrant ces effectifs, et il est d'autant plus grand que la maladie, dans la population étudiée, est plus rare : dans le cas du cancer du poumon, il suffit de quelques centaines de sujets dans chacun des groupes malade et témoin pour obtenir la même précision qu'avec 200 000 sujets d'un échantillon aléatoire.

En réalité il arrive souvent, dans les enquêtes de ce genre, que les témoins soient plus faciles à recruter que les malades, de sorte qu'on préfère en réunir un plus grand nombre ($n_{*0} > n_{*1}$). On se souviendra toutefois qu'il n'est pas opportun d'aller trop loin dans cette voie; l'expression $\frac{1}{n_{*0}} + \frac{1}{n_{*1}}$ ne diminue plus guère, pour n_{*1} donné, quand n_{*0} devient grand : c'est ainsi qu'entre la valeur atteinte pour $n_{*0} = 3n_{*1}$ (soit $\frac{4}{3}\frac{1}{n_{*1}}$) et pour n_{*0} infini (soit $\frac{1}{n_{*1}}$) la diminution de variance ne compense guère la difficulté de recrutement.

Un cas extrême du type 3 est celui où on constitue seulement le groupe malade, le groupe témoin s'identifiant à la population générale - ceci n'étant possible que si la fréquence d'exposition au facteur est connue pour celle-ci, et si la maladie est suffisamment rare pour qu'on puisse confondre population non malade et population générale : on a comparé par exemple aux données de la population générale la situation de famille observée sur un groupe de 1 200 femmes atteintes de cancer du col utérin (48), ou la fréquence de la mortalité par cancer du sein dans l'ascendance féminine d'un groupe de malades atteintes de ce même cancer (46). Ces comparaisons sont naturellement faites à âge égal, éventuellement à milieu social égal, etc. par les corrections exposées plus loin.

D'une manière générale, dans les enquêtes du type 3, et surtout lorsqu'on craint des biais dans l'échantillonnage du groupe témoin, on devra tenir compte des multiples facteurs d'échantillonnage, pour des raisons qui ont été détaillées dans l'exemple de "l'enquête rétrospective".

d) CONCLUSION.

Le type 1, avec son échantillonnage représentatif de la population étudiée, est très coûteux en nombre de sujets.

Le type 2 permet de réduire ce nombre, ceci d'autant plus que l'exposition au facteur est une éventualité plus rare.

Le type 3 permet une réduction du même genre, d'autant plus considérable que la maladie est plus rare.

Il va de soi qu'en contre-partie on ne saurait attendre autant des types 2 et 3 que du type 1 : ils ne peuvent donner que des conclusions moins étendues et d'une valeur plus discutable; c'est ce que précisera le chapitre suivant.

Tableau 1

Echantillon (effectifs)

		malades		
		non	oui	total
exposés	non	n_{00}	n_{01}	n_{0*}
	oui	n_{10}	n_{11}	n_{1*}
Total		n_{*0}	n_{*1}	n

Tableau 2

Population générale (proportions)

	Malades non	Malades oui	Total	Proportion de sujets malades dans le groupe
exposés non	p_{00}	p_{01}	$p_{0*} = 1 - x$	$\frac{p_{01}}{p_{0*}} = m_0$
exposés oui	p_{10}	p_{11}	$p_{1*} = x$	$\frac{p_{11}}{p_{1*}} = m_1$
Total	$p_{*0} = 1 - m$	$p_{*1} = m$	1	
Proportion de sujets exposés dans le groupe	$\frac{p_{10}}{p_{*0}} = x_0$	$\frac{p_{11}}{p_{*1}} = x_1$		

II - TEST ET MESURE DU ROLE ETIOLOGIQUE DE L'EXPOSITION AU FACTEUR -

a) MESURE DU ROLE ETIOLOGIQUE DANS LA POPULATION ETUDIEE, SUPPOSEE HOMOGENE.

Nous nous plaçons dans le cas du modèle simplifié, décrit plus haut, et représenté au tableau 2, où on envisage 4 catégories de sujets, exposés ou non exposés, malades ou non.

Nous supposons en outre, pour commencer, que la population étudiée est homogène pour les facteurs essentiels énumérés plus haut, tels que : âge, sexe, niveau social...

Indépendamment des proportions ou probabilités p_{00}, p_{01}, p_{10}, p_{11}, qui définissent entièrement la situation, nous avons fait figurer simultanément au tableau 2 quelques combinaisons de ces probabilités qui, pour simplifier, seront désignées par des symboles plus parlants : m_0, m_1, et m , les probabilités de maladie chez les sujets non exposés, exposés, et globalement ; x_0, x_1, et x les proportions de sujets exposés au facteur x parmi les sujets indemnes, malades, et globalement.

Si l'exposition au facteur n'intervient pas dans l'étiologie de la maladie m, les probabilités p_{01} et p_{11} sont proportionnelles à p_{00} et p_{10} , ou encore les probabilités m_0 et m_1 sont égales (ainsi d'ailleurs que

les proportions de sujets exposés x_0 et x_1). Si elle intervient, il n'en est pas ainsi, les probabilités m_0 et m_1 par exemple sont différentes, en principe dans le sens $m_1 > m_0$.

Si l'exposition au facteur a un rôle étiologique (cette locution ne supposant pas qu'il s'agisse d'une relation causale) on peut se proposer de traduire ce rôle quantitativement.

Il est d'abord certain que le rôle du facteur x est d'autant plus important que le tableau 2 s'écarte davantage du modèle de l'indépendance, c'est-à-dire par exemple que m_1 s'écarte davantage de m_0. On pourra donc mesurer ce rôle par une expression indiquant l'écart entre m_1 et m_0.

Hammond et Horn, dans l'enquête prospective sur la mortalité en relation avec l'usage du tabac (32), ont utilisé, pour une cause de décès donnée, par exemple le cancer du poumon, le rapport $\frac{m_1}{m_0}$, qui mesure la surmortalité des fumeurs. Berkson (4) pense qu'il vaudrait mieux utiliser la différence $(m_1 - m_0)$. De toute manière, aucune fonction de m_1 et m_0 ne peut à elle seule résumer la situation définie par les 2 données m_1 et m_0; il est clair que pour un rapport donné la différence peut être très variable, et inversement. Sheps (61) souligne qu'il est plus intéressant de former telle ou telle fonction de m_1 et m_0 qui ait un sens concret dans un modèle donné. Il propose notamment de faire intervenir la mortalité par cancer du poumon liée en propre à l'usage de la cigarette, soit m_x, et d'écrire la mortalité chez les fumeurs sous la forme :

$$m_1 = m_0 + m_x - m_0 m_x \qquad (1)$$

C'est là un modèle particulièrement simple, car aux conventions déjà adoptées plus haut (on n'envisage pas que les sujets non exposés au facteur x puissent être exposés ou non à d'autres facteurs, ce qui conduirait à un schéma à plus de 2 dimensions), on ajoute une hypothèse supplémentaire : les sujets, qu'ils soient exposés ou non exposés au facteur x (tabac), auraient par ailleurs la même probabilité de décès par cancer broncho-pulmonaire pour les "autres causes". Si on adopte ce schéma en première approximation, de (1) on tire :

$$m_x = \frac{m_1 - m_0}{1 - m_0} \qquad (2)$$

Cette fonction de m_1 et m_0 a un sens concret, puisqu'elle mesure la mortalité liée en propre à l'exposition au facteur x, ou encore mortalité qu'on observerait en l'absence des autres causes de cancer

broncho-pulmonaire; c'est surtout dans le cas de la relation causale que cette expression est intéressante : m mesure alors l'effet propre du facteur x.

On notera que $1 - m_x = \frac{1 - m_1}{1 - m_0}$; ce dernier rapport, qui prend ainsi un sens concret, est le rapport des survies des groupes exposés et non exposés, de sorte que le rapport des survies devient plus intéressant que le rapport des mortalités.

Il va de soi que m_x, pas plus qu'une autre fonction, ne résume m_1 et m_0, et qu'il faut une deuxième information pour définir le couple (m_1, m_0); celle-ci peut être m_0, le couple (m_0, m_x) ayant une valeur plus concrète que le couple (m_0, m_1), puisqu'il exprime le risque en l'absence du facteur, et le risque lié en propre à l'exposition au facteur.

Enfin le couple (m_0, m_x) ne suffit pas encore à résumer la situation décrite par le tableau 2; celui-ci est défini par 4 probabilités p_{00}, p_{01}, p_{10}, p_{11}, dont la somme est 1, donc par 3 données indépendantes. On peut alors adjoindre au couple (m_0, m_x) une troisième donnée, par exemple la fréquence de l'exposition au facteur, soit x. La situation serait alors ainsi résumée :

m_0 : probabilité de maladie en l'absence d'exposition au facteur;

m_x : probabilité de maladie pour un sujet exposé, en l'absence d'autres causes de la maladie, ou effet propre du facteur dans l'hypothèse causale;

x : fréquence de l'exposition au facteur.

On peut naturellement préférer un autre groupe de 3 indices. Il reste que, de toute manière, le rôle étiologique d'un facteur ne saurait être mesuré par un seul indice : c'est là un résultat commun à tout problème de liaison entre 2 variables aléatoires dichotomiques et qu'on rencontre sous une forme similaire quand on veut mesurer le rôle d'un critère en matière de pronostic ou de diagnostic.

Un indice intéressant est la proportion de cas dus au facteur (proportion de cancers du poumon dus à l'usage du tabac), soit x_p. C'est :

$$x_p = \frac{xm_1 - xm_0}{(1 - x)m_0 + xm_1} = \frac{x(m_1 - m_0)}{m_0 + x(m_1 - m_0)} \qquad (3)$$

ou, en fonction de m_0, m_x, de x,

$$x_p = \frac{x(1 - m_0)m_x}{m_0 + x(1 - m_0)m_x} \tag{4}$$

b) TEST ET MESURE DU ROLE ETIOLOGIQUE D'APRES L'ECHANTILLON REPRESENTATIF D'UNE POPULATION HOMOGENE.

Pour commencer, nous supposons ici l'échantillonnage correct, c'est-à-dire donnant un échantillon représentatif de la population étudiée dans le type 1, deux groupes représentatifs des catégories exposée et non exposée dans le type 2, malade et témoin dans le type 3.

Nous supposons encore qu'il s'agit, dans chacun de ces cas, d'une population homogène en ce qui concerne les caractéristiques essentielles énumérées dès l'introduction de cet exposé, c'est-à-dire de sexe donné, d'âge donné, éventuellement de niveau social ou de milieu d'habitation donné ...

Il s'agit, d'après l'échantillon observé, d'éprouver puis d'estimer le rôle étiologique de l'exposition au facteur.

Le TYPE 1 permet de connaître complètement le rôle étiologique du facteur : on éprouve d'abord ce rôle par jugement sur l'échantillon du tableau 1, à l'aide d'un test classique (X^2 sur le tableau 2 × 2); il est possible ensuite d'estimer m_0 par $\frac{n_{01}}{n_{0*}}$, m_1 par $\frac{n_{11}}{n_{1*}}$, et x par $\frac{n_{1*}}{n}$.

Dans le TYPE 2, on a encore des estimations valables de m_0 et m_1, et leur comparaison permet d'éprouver le rôle étiologique du facteur. Cette comparaison de proportions se ramène, ici encore, à un test de X^2 sur le tableau 2 × 2. La mesure du rôle étiologique ne saurait par contre être complète : on a des estimations de m_0 et m_1 comme ci-dessus. Mais la fréquence x de l'exposition n'est pas connue, puisqu'on a choisi arbitrairement les effectifs des groupes exposé et non exposé. Ce mode d'enquête ne permet donc que d'évaluer l'effet du facteur, mais pas sa fréquence. (Notons qu'il est parfois possible de connaître celle-ci par ailleurs, à l'aide de données statistiques générales).

Dans le TYPE 3, une difficulté se présente dès le test d'association : on ne peut pas comparer m_1 et m_0, car on ne dispose pas de leurs estimations du fait qu'on a choisi arbitrairement les effectifs des groupes malade et témoin. Par contre on a des estimations correctes de x_1 et x_0, qu'on peut comparer par un test de signification, qui est ici encore un X^2 sur le tableau 2 × 2. Or il est visible que ce test permet d'éprouver le rôle étiologique du facteur. Si les désigna-

tions "malade" ou "non malade" des tableaux 1 et 2 désignent des sujets présentant la maladie pendant l'époque de l'enquête(1), le test d'association est réversible : si $x_1 > x_0$, on a aussi $m_1 > m_0$, c'est-à-dire une fréquence des cas de maladie, dénombrables pendant un intervalle de temps donné, plus élevée dans le groupe exposé, ce qui indique le rôle étiologique du facteur.

La mesure de ce rôle est malaisée : on peut estimer seulement x_0 et x_1, ce qui donne comme dans le type 2 deux indices au lieu de 3; mais ces indices ne sont guère intéressants, et on ignore m_0, m_1, et x, - à moins naturellement que la fréquence de la maladie dans la population générale ne soit par ailleurs connue par des données statistiques, auquel cas, disposant de 3 données, on peut connaître complètement le rôle du facteur.

Toutefois, lorsque la fréquence de la maladie, sans être connue, est faible, on peut tirer de l'enquête des renseignements étiologiques plus intéressants; si on suppose la maladie rare, tant pour le groupe exposé que pour le groupe non exposé, on a en effet :

$$m_0 = \frac{p_{01}}{p_{00} + p_{01}} \# \frac{p_{01}}{p_{00}}$$

$$m_1 = \frac{p_{11}}{p_{10} + p_{11}} \# \frac{p_{11}}{p_{10}}$$

et

$$\frac{m_1}{m_0} = \frac{p_{11}/p_{10}}{p_{01}/p_{00}} = \frac{p_{11}}{p_{01}} \times \frac{p_{00}}{p_{01}} \qquad (5)$$

expression qui peut être estimée, à partir des données, par

$$r = \frac{n_{11}}{n_{01}} \times \frac{n_{00}}{n_{10}} \qquad (6)$$

Le rapport $\frac{m_1}{m_0}$ a été appelé risque relatif par Cornfield (15), qui en a donné l'estimation par la formule (6), ainsi que les limites de confiance. Ce risque relatif r mesure le rapport entre les proportions de sujets présentant la maladie donnée, pendant un intervalle de temps déterminé, chez les sujets exposés et non exposés. Dans le cas du cancer broncho-pulmonaire par exemple, le risque relatif des fumeurs

----

(1) Il s'agit donc, pour reprendre la distinction définie plus haut, des cas existant à un moment donné (en anglais prevalence).

par rapport à celui des non fumeurs, est de l'ordre de 10.

Le risque relatif reste naturellement soumis aux limitations indiquées plus haut pour le rapport $\frac{m_1}{m_0}$; il ne saurait à lui seul résumer m_1 et m_0, et 2 situations étiologiques caractérisées, l'une par $m_0 = 1/1\ 000$, $m_1 = 10/1\ 000$, l'autre par des proportions 10 fois plus élevées, donnent le même risque relatif $r = 10$ alors que $(m_1 - m_0)$ par exemple est très différente. Mais le risque relatif a pour lui de pouvoir être estimé à partir des données, ce qui n'est le cas ni pour $(m_1 - m_0)$, ni pour m_x...

La place du facteur x dans l'étiologie peut également, - toujours dans la même hypothèse de maladie rare et dans le cas du modèle décrit plus haut - être connue. La proportion de cas dus à la maladie étant d'après (3) :

$$x_p = \frac{x(m_1 - m_0)}{(1 - x)m_0 + xm_1}$$

on a, pour une maladie rare,

$$x \# p_{10}, \qquad 1 - x \# p_{00}, \qquad m_0 \# \frac{p_{01}}{p_{00}}, \qquad m_1 \# \frac{p_{11}}{p_{10}}$$

de sorte que :

$$x_p \# \frac{p_{11} - \frac{p_{10}p_{01}}{p_{00}}}{p_{01} + p_{11}} = \frac{1 - \frac{p_{10}p_{01}}{p_{00}p_{11}}}{\frac{p_{01}}{p_{11}} + 1} \tag{7}$$

expression dépendant seulement de $\frac{p_{10}}{p_{00}}$ et $\frac{p_{01}}{p_{11}}$, qu'on peut estimer d'après l'échantillon.

En faisant intervenir les proportions de sujets exposés, dans les groupes malade

$$x_1 = \frac{p_{11}}{p_{01} + p_{11}}$$

et non malade

$$x_0 = \frac{p_{10}}{p_{00} + p_{10}}$$

on peut exprimer x_p sous les formes :

$$x_p = \frac{x_1 - x_0}{1 - x_0} \qquad (8)$$

ou

$$x_p = \frac{x_0(r - 1)}{x_0(r - 1) + 1} \qquad (9)$$

proposée par Levin (39).

c) TEST ET MESURE DU ROLE ETIOLOGIQUE, DANS LE CAS D'UNE POPULATION HETEROGENE (ELIMINATION DE L'INFLUENCE DES TIERS FACTEURS).

Il arrive le plus souvent que la population étudiée soit hétérogène au regard des facteurs déclarés essentiels, tels que sexe, âge, milieu d'habitation ... Il s'agit alors, d'après un échantillon reflétant cette hétérogénéité, d'éprouver puis de mesurer le rôle étiologique de l'exposition au facteur.

On peut diviser chacun des facteurs essentiels en classes, par exemple : 5 tranches d'âge, 4 niveaux sociaux, 3 milieux d'habitation (grande ville, petite ville, campagne). Les diverses combinaisons de ces classes constituent c "cellules" (ici 5 × 4 × 3 = 60 cellules).

Chacune de ces cellules est homogène.

Une première solution du problème consiste à étudier séparément chaque cellule, par les procédés indiqués précédemment, autrement dit à subdiviser l'enquête en c sous-enquêtes; c'est la seule solution réellement correcte; elle permet d'observer éventuellement des résultats différents selon les cellules.

Cependant, dès que le nombre des cellules est élevé, les effectifs y deviennent trop faibles pour que ce procédé soit applicable.

On est alors conduit à étudier simultanément les c sous-enquêtes par une analyse d'ensemble, en supposant réalisées certaines hypothèses d'identité des résultats d'une cellule à l'autre. Cette analyse vise ainsi à corriger l'hétérogénéité de la population, en indiquant ce que serait le rôle étiologique de l'exposition au facteur à âge, niveau social, et milieu d'habitation donnés, donc à éliminer l'influence de ces tiers facteurs.

La première partie de cette analyse est le test du rôle étiologique. L'hypothèse faite pour permettre une étude simultanée des diverses cellules est que, si l'exposition au facteur joue un rôle dans l'étiolo-

gie, ceci doit être vrai dans toutes les cellules, et l'hypothèse nulle est l'absence de rôle étiologique dans chacune des cellules.

On est alors ramené à éprouver, par un test unique, l'absence de liaison dans un ensemble de tableaux de contingence 2 × 2.

Ce problème est justiciable de plusieurs solutions (voir notamment 13, 42) :

1/ On peut utiliser la somme des X^2, avec la somme des degrés de liberté (c). Ce test présente un inconvénient : il ne tient pas compte du signe de la différence dans chaque cellule.

2/ On peut comparer à O la moyenne des X par l'écart-réduit

$$\frac{(X \text{ moyen}) - O}{1/\sqrt{c}} = X\sqrt{c}$$

Ce test est en général meilleur que le précédent, mais il a encore l'inconvénient d'attribuer un même poids aux cellules, quel que soit leur effectif.

3/ Une meilleure solution consiste à donner des poids aux diverses cellules. Adoptons les désignations ci-dessous pour la cellule i (dans le cas du type 3; s'il s'agit du type 1ou 2 on intervertira les termes "malade" et "exposé").

	Malades	Non malades	Différence
Effectif	n_i	n'_i	
Proportion de sujets exposés	p_i	p'_i	d_i
Proportion de sujets non exposés	q_i	q'_i	

Le test de l'égalité à O de l'ensemble des d_i peut être effectué en comparant à O une combinaison pondérée $\Sigma\, a_i d_i$, où on calculera les a_i de façon à obtenir le test le plus puissant.

Si on désigne par P_i et Q_i les proportions dans l'ensemble de la cellule, et si on pose $\frac{1}{w_i} = \frac{1}{n_i} + \frac{1}{n'_i}$, Cochran (13) propose comme solution le test :

$$X^2 = \frac{(\Sigma\, w_i\, d_i)^2}{\Sigma w_i\, P_i\, Q_i}$$

avec un degré de liberté.

Mantel et Haenszel (49) proposent un test très voisin, modifié pour tenir compte de la correction de continuité.

4/ On peut également, dans chaque cellule, calculer les effectifs théoriques des 4 cases dans l'hypothèse d'indépendance pour cette cellule. On somme ensuite les effectifs des cases homologues de toutes les cellules; aux 4 effectifs théoriques ainsi obtenus on compare les 4 effectifs observés sur l'échantillon total, par un X^2 à 1 degré de liberté (6). L'intérêt de ce test, par rapport aux précédents, est qu'il est facilement généralisable au cas où l'exposition au facteur x comporte plus de 2 classes, à condition de prendre le nombre de degrés de liberté voulu.

D'autres tests ont également été proposés. En fait, la neutralisation de variables plus ou moins nombreuses dans la comparaison de 2 groupes est un problème très général, mais il est si important dans le cas des enquêtes médicales qu'il constitue l'élément principal de leur analyse. Ceci explique la variété des tests utilisés.

La plupart des procédés indiqués rappellent la "standardisation par âge" utilisée par les démographes pour comparer "à âge égal" 2 populations dont la distribution d'âge est différente. Aussi sont-ils communément appelés standardisation par âge, situation sociale, milieu d'habitation, etc.

La standardisation est appliquée d'une manière relativement empirique, tant par le choix des tiers facteurs retenus (qui peut être restreint ou étendu) que par leur division en classes, et la constitution finale des cellules : il arrive qu'on standardise par rapport à chaque facteur isolément, ou par rapport à des groupes de deux, plutôt que de procéder à une standardisation d'ensemble conduisant à un grand nombre de cellules d'effectif très faible. Il peut arriver également qu'on fasse une étude séparée par sexe, avec pour chaque étude une standardisation pour les autres facteurs. Le choix entre les diverses voies d'approche est une affaire d'opportunité.

Les tests qui viennent d'être décrits, pour complexes qu'ils soient, ne représentent qu'un premier pas : l'épreuve d'association. Si l'exposition au facteur s'avère jouer un rôle étiologique, il reste à le mesurer.

Cette mesure peut être faite dans chaque cellule, mais l'élaboration d'une mesure unique, englobant les résultats de toutes les cellules, soulève des difficultés, car elle n'a de sens que si on suppose une comparabilité de toutes les cellules, qui est rarement vérifiée : dans le cas, par exemple, des enquêtes rétrospectives de type 3, une

combinaison pondérée des risques relatifs ne paraît intéressante que si l'espérance mathématique de ces risques est la même dans toutes les cellules, hypothèse peu vraisemblable. Diverses combinaisons pondérées, de nature empirique, ont été proposées (49), mais leur difficulté d'interprétation ne fait que souligner les limitations d'emploi du risque relatif.

Un mot doit être dit enfin des enquêtes avec appariement; l'appariement peut être utilisé, dans les enquêtes de type 2 et 3, pour rendre comparables, vis-à-vis de certains facteurs, les groupes exposé et non exposé ou malade et non malade. C'est donc une méthode visant, dès le stade de l'échantillonnage, à corriger l'hétérogénéité de l'influence de tiers facteurs.

Cependant l'appariement ne remplit complètement sa fonction que si l'analyse statistique en tient compte. On peut utiliser les tests classiques pour la comparaison de deux proportions dans des séries de sujets appariés (voir notamment 49). Ces méthodes sont généralisables au cas où l'exposition au facteur comporte plus de 2 classes (41).

Le gain de précision conféré par l'appariement n'est intéressant que si la variable d'appariement est fortement liée à l'exposition au facteur (12).

d) VALIDITE DES RESULTATS.

Les perfectionnements mathématiques apportés à l'analyse statistique ne doivent pas faire perdre de vue diverses erreurs portant sur les données de base, et qui peuvent retirer toute valeur aux conclusions.

Il s'agit d'étudier l'association entre 2 variables x et m.

Ces variables sont d'abord passibles d'une erreur d'appréciation : on peut classer un sujet comme fumeur alors qu'il ne l'est pas, et inversement; comme atteint de cancer du poumon alors qu'il en est indemne, et inversement; de telles erreurs sont inévitables (ne serait-ce que parce qu'un sujet témoin souffre peut-être d'un cancer encore inapparent), mais il importe d'en distinguer 2 catégories :

- les erreurs portant sur une des variables sans relation avec l'autre ne sont pas graves : classer quelques sujets dans un groupe au lieu de l'autre revient à atténuer l'écart entre ces groupes et à diminuer la puissance du test, mais ne risque pas d'entraîner des conclusions erronées;

- beaucoup plus graves sont, par contre, les erreurs in-

fluencées par la liaison à étudier : si, parce que le sujet est atteint d'un cancer du poumon, lui-même ou l'enquêteur qui l'interroge exagèrent sa consommation de tabac, si inversement le médecin fait intervenir dans les éléments de son diagnostic de cancer une consommation de tabac élevée, alors on risque d'observer une association reflétant uniquement l'idée préconçue. Or la subjectivité des réponses est souvent manifeste : les malades atteints d'un cancer du pharynx se remémorent ou insistent davantage sur les maux de gorge antérieurs, les femmes atteintes d'un cancer du sein ont tendance à exagérer la fréquence des douleurs mammaires dans leur passé, ou des cancers du sein dans leur famille. D'une manière générale, la comparabilité des interrogatoires entre sujets malades et témoins, ou exposés et non exposés, est une des difficultés majeures de l'enquête : comment obtenir qu'une mère, dont l'enfant est mort de leucémie, réponde à l'interrogatoire de la même façon qu'une mère témoin ?

Aussi l'élimination de ce type d'erreur doit-elle être recherchée par tous les moyens.

Il faut d'abord obtenir un diagnostic indépendant du facteur x, ce qui est facile si le diagnostic repose sur des éléments objectifs, par exemple l'histologie pour un cancer.

Il faut ensuite obtenir, pour le facteur x, des informations indépendantes du diagnostic; c'est ici que l'enquête "prospective" décrite plus haut offre des garanties supérieures à toute autre, puisque l'interrogatoire a lieu à un moment où la maladie n'est pas encore déclarée. Dans les autres modes d'enquête, lorsque la maladie est déjà déclarée au moment de l'interrogatoire, l'ignorance du diagnostic par le malade, l'enquêteur, ou les deux, doit être recherchée dans toute la mesure du possible : on interrogera par exemple comme malades et témoins des sujets consultant pour une tumeur qui n'est cataloguée qu'ultérieurement comme maligne ou bénigne; Doll et Hill (22) ont ainsi apporté un argument important en signalant que la proportion de fumeurs était normale chez des sujets étiquetés "cancer du poumon" au moment de l'interrogatoire, et dont le cancer a été infirmé ultérieurement. Un autre argument important est que certains types histologiques seulement sont liés à l'usage du tabac, et pas d'autres, alors que le type histologique n'est pas connu au moment de l'interrogatoire.

Les considérations précédentes avaient trait aux erreurs de mesure; des réserves analogues doivent être énoncées pour les erreurs d'échantillonnage, qu'on doit également subdiviser en 2 catégories :

- les erreurs d'échantillonnage portant sur une des variables ne sont pas trop graves : si dans une enquête du type 1 l'échantil-

lon observé n'a pas tout à fait le même milieu social que la population étudiée, la consommation de tabac sera peut-être faussée, mais les méfaits éventuels de cette consommation le seront sans doute peu; si, dans les enquêtes de types 2 et 3 l'échantillonnage des groupes à comparer diffère quelque peu, on pourra corriger ces différences par une standardisation.

- beaucoup plus graves sont par contre les erreurs d'échantillonnage portant sur l'association même des 2 variables.

Les enquêtes du type 1 présentent à cet égard une certaine sécurité. Berkson (2) a, il est vrai, imaginé une cause d'erreur possible dans les enquêtes prospectives sur les fumeurs ; cependant un tel biais reste minime (35).

Il n'en est pas de même dans les enquêtes du type 2, et surtout du type 3; un premier exemple d'erreur a été signalé par Berkson (1) : c'est le cas où on étudie l'association entre une maladie m et une autre maladie, jouant le rôle du facteur x, parmi les malades se présentant à l'hôpital. Les sujets souffrant de la maladie m ont une certaine propension à se rendre à l'hôpital. S'ils souffrent en outre de la maladie x, cette propension est plus élevée, de sorte que l'échantillon hospitalier de malades (m) montrera une proportion trop élevée de sujets souffrant de la maladie x. Chez les témoins - qui sont les malades souffrant de diverses maladies - ce biais existe également, plus fortement ou moins fortement que pour la maladie m, selon le cas ; la comparaison des 2 groupes peut alors faire apparaître des différences purement artificielles.

La même situation se présente lorsqu'on étudie l'association entre 2 maladies ou signes morbides dans une série d'autopsies : par exemple entre nodules tuberculeux et cancer. Les 4 combinaisons, avec et sans cancer, avec et sans nodules, sont, chez des sujets décédés, différentes de celles qui existent dans la population générale, en raison de leurs taux de mortalité différents; on peut admettre que, chez les sujets non cancéreux, la présence de nodules augmente la mortalité, tandis que cet effet est négligeable chez les cancéreux : ainsi apparaîtra illusoirement chez les décédés une association négative, entre nodules tuberculeux et cancer, qui n'existe pas dans la population générale des vivants (2, 47).

De tels biais soulignent une limitation de ce genre d'enquête où malades et témoins se recrutent d'eux-mêmes par leur venue dans l'échantillon (par la décision de consulter, par la mort, etc.) : c'est qu'il n'est pas possible d'étudier le rôle étiologique d'un facteur influençant le recrutement, ou du moins l'influençant inégalement pour les malades et les témoins.

Il faut bien le dire : le statisticien, habitué à constituer un échantillon par des procédés classiques de tirage au sort, risque d'être surpris, voire choqué, en découvrant que dans la plupart des enquêtes médicales on laisse aux sujets la responsabilité de l'auto-recrutement. Pour une maladie donnée, le fait d'aller à l'hôpital occasionne déjà une première sélection, dépendant de facteurs sociaux et psychologiques. Les sujets présents un jour donné à l'hôpital constituent une nouvelle sélection, un malade ayant d'autant plus de chances d'être présent que sa durée d'hospitalisation est plus longue (57). Dans le même ordre d'idées, l'échantillon de malades vivants un jour donné constitue également une sélection renforçant la proportion de malades à survie longue (53). La notion de représentativité fait trop souvent place à la notion de commodité, et il arrive, comme le fait remarquer Dorn dans une mise au point récente (27), qu'une enquête du type 3 vise à comparer "deux échantillons sans spécification provenant par une méthode d'échantillonnage inconnue d'une population non identifiée".

Ceci ne doit pas être considéré comme une condamnation des enquêtes du type 3, qui restent le seul moyen facilement réalisable de suggérer des facteurs étiologiques; mais leurs conclusions doivent être accueillies avec réserve, et soumises, lorsque l'enjeu en vaut la peine, à la confirmation d'enquêtes du type 1, plus rigoureuses mais infiniment plus difficiles à entreprendre.

e) CONCLUSION.

L'analyse du rôle étiologique d'un facteur peut être complète dans le type 1; elle est nécessairement incomplète dans le type 2 et surtout dans le type 3.

La validité des résultats ne peut d'autre part être garantie que si l'on a pu éviter des erreurs de mesure et des erreurs d'échantillonnage portant précisément sur la liaison à étudier; à cet égard on peut obtenir une relative sécurité avec les enquêtes du type 1, tandis que les biais sont plus difficilement évitables dans le type 2 et surtout dans le type 3.

III - L'INTERPRETATION CAUSALE -

Après avoir évité les biais et pièges de tout ordre, éliminé le rôle de quelques "tiers facteurs" essentiels (sexe, âge ...), on conclut au rôle étiologique de l'exposition au facteur x. Peut-on interpréter ce rôle en termes de causalité ?

Nous avons souligné dès le départ l'impuissance fondamentale à cet égard de l'enquête d'observation; dans une expérimentation, on peut

exposer ou non au facteur 2 groupes comparables à tout point de vue, de sorte que toute différence revêt d'emblée une signification causale; dans l'observation l'exposition au facteur est déterminée aléatoirement, en liaison avec d'autres facteurs x_1, x_2, ... parmi lesquels peut se trouver la vraie cause : les fumeurs étant plus souvent des citadins et des buveurs de café, la vraie cause du cancer du poumon ne serait-elle pas l'abus du café, ou l'atmosphère polluée des villes ? En outre, dans une expérience, on peut souvent maintenir la comparabilité entre les groupes exposé et non exposé après l'intervention du facteur, tandis que dans les conditions spontanées ces 2 groupes peuvent se différencier systématiquement : les fumeurs, sujets au catarrhe, devront peut-être davantage se faire radiographier : si les rayons X étaient alors la cause du cancer broncho-pulmonaire, l'usage du tabac serait certes un facteur causal, mais par une voie indirecte dont la signification serait très différente de la causalité directe.

De fait, bien des facteurs apparus, au cours d'une enquête, comme associés à l'apparition d'une maladie, sont sans action causale réelle : le niveau social pour le cancer de l'estomac (14), la presbytie précoce dans le cas de la maladie coronarienne (8), entrent sans doute dans cette catégorie.

Par contre, dans d'autres cas, comme celui du tabac pour le cancer du poumon, il est possible de justifier une forte présomption de causalité.

Il est d'abord possible d'argumenter contre l'objection du "tiers facteur". Si la "vraie cause" du cancer broncho-pulmonaire est un facteur lié à l'usage du tabac, par exemple l'abus du café, on doit alors observer qu'à consommation de café donnée le rôle du facteur tabac disparaît ; cette étude du rôle du tabac à niveau égal pour différents autres facteurs peut être effectuée, ceci par les divers procédés envisagés plus haut (standardisation). L'élimination des "tiers facteurs" entreprise déjà pour certains facteurs essentiels (sexe, âge ...) dans l'épreuve du rôle étiologique, peut être poursuivie avec plus de détails pour toute une série de facteurs liés à l'exposition au facteur x (ceci est nécessaire quel que soit le type d'enquête). Ce travail a été fait, dans le cas du cancer broncho-pulmonaire, et on a observé que la prise en considération de plusieurs dizaines de facteurs ne permettait en aucun cas d'"innocenter" le tabac (20).

Sans doute cette méthode d'exploration est-elle soumise à une sérieuse limitation : elle ne permet d'étudier que les facteurs prévus dans l'interrogatoire des malades; or la "vraie" cause peut être insoupçonnée. Mais ici intervient un argument d'ordre quantitatif : la liaison

entre l'usage du tabac et l'apparition du cancer broncho-pulmonaire étant très forte, il est facile de montrer qu'elle ne peut être "expliquée" par un tiers facteur, que si celui-ci est à la fois très lié à l'apparition de ce cancer et à l'usage du tabac (16); il est peu probable qu'un tel facteur ait échappé aux nombreuses investigations effectuées. Il n'est pas suffisant, pour invalider l'hypothèse causale, de déceler par exemple un effet de l'hérédité dans l'habitude de fumer : il faudrait encore que la constitution génétique fût fortement liée à cette habitude (ce qui n'est pas le cas), et à l'apparition du cancer broncho-pulmonaire (ce qui n'a pas été signalé). D'une manière générale, plus l'association est forte, entre l'exposition au facteur et l'apparition de la maladie, et plus la présomption causale est solide.

Cependant l'élimination de tous les tiers facteurs n'est pas concevable (le fût-elle qu'elle n'apporterait d'ailleurs pas la certitude : à la limite, si les cancéreux et les témoins ne différaient que par l'usage du tabac, on pourrait supposer que c'est le cancer qui conduit les sujets à fumer ...).

C'est pourquoi d'autres arguments, de divers ordres, doivent être recherchés. Indiquons que, dans le cas du facteur tabac, - indépendamment des confirmations obtenues en laboratoire, in vitro et in vivo, dont la contribution est toujours essentielle -, on a observé les relations suivantes :

- la probabilité de cancer broncho-pulmonaire est plus élevée chez le fumeur que chez le non fumeur, d'autant plus qu'il fume davantage, selon une loi proportionnelle ; elle diminue si le sujet s'est arrêté de fumer, et d'autant plus qu'il s'est arrêté plus tôt;

- on rencontre chez les fumeurs une proportion exagérée de cancers de la cavité buccale, du pharynx, du larynx, de l'œsophage, et de la vessie - c'est-à-dire de toutes les localisations directement exposées à la fumée ou à ses dérivés immédiats - et une proportion normale des autres cancers;

- la probabilité de cancer est augmentée par le fait de respirer la fumée lorsqu'il s'agit du cancer des bronches ou du larynx, elle n'est pas augmentée pour les autres cancers des voies aéro-digestives supérieures.

La convergence parfaite de ces arguments dans le sens de la relation causale ne peut manquer de frapper. Si l'usage du tabac n'est pas la "vraie" cause, il faut qu'il accompagne celle-ci bien fidèlement : présent quand elle est présente, absent si elle disparaît, faible ou fort à sa mesure. L'hypothèse d'un facteur aussi "mimétique" ne saurait être écartée avec certitude, mais elle fait penser à la phrase

de l'humoristé : on a découvert l'auteur des pièces de Shakespeare; c'est un homme qui vivait à la même époque, dans le même village, et qui portait le même nom que lui...

Ce n'est pas, cependant, sur l'étiologie du cancer broncho-pulmonaire qu'il serait équitable de terminer : si on a pu, dans ce cas, après plus de 25 enquêtes, parvenir à la quasi-certitude, c'est qu'il s'agit d'un cas facile : la probabilité de cancer broncho-pulmonaire est extrêmement faible chez un non fumeur, elle est dix fois plus élevée si le sujet fume, et l'usage du tabac est très répandu; l'exposition au facteur joue donc dans l'étiologie de cette maladie un rôle considérable.

Mais dans beaucoup d'enquêtes, la situation se présente moins favorablement : alors un travail ardu d'interrogatoire, une analyse statistique complexe pour tenir compte de multiples variables, ne conduisent, en ce qui concerne la relation causale, qu'à des conclusions incertaines : seule est responsable de cette faible rentabilité la complexité même du sujet.

BIBLIOGRAPHIE

Les références notées ++ portent sur la méthodologie des enquêtes.
Les références notées + correspondent à des enquêtes où sont exposés ou discutés certains points de la méthodologie.

++ (1) BERKSON J. - Limitations of the application of fourfold table analysis to hospital data. Biometrics bulletin 2, N°3 : 47-53, 1946.

++ (2) BERKSON J. - The statistical study of association between smoking and lung cancer. Proceed. of the Staff Meetings of the Mayo Clinic 30, N°15, 1955.

(3) BERKSON J., COMFORT M.W., BUTT H.R. - Occurence of gastric cancer in persons with achlorhydria and with pernicious anemia. Proceed. of the Staff Meetings of the Mayo Clinic 31 : 583-596, 1956.

++ (4) BERKSON J. - The statistical investigation of smoking and cancer of the lung. Proceed. of the Staff Meetings of the Mayo Clinic 34 : 206-224 a, 1959.

(5) BILLINGTON B.P. - Gastric cancer - Relationships between a-b-o blood-groups, site, and epidemiology. The Lancet : 859-862, 1956.

+ (6) BOYD J.T., DOLL R. - Gastro-intestinal cancer and the use of liquid paraffin. Brit. J. Cancer 8 : 231-237, 1954.

(7) BRESLOW L., HOAGLIN L., RASMUSSEN G., ABRAMS H.K. - Occupations and smoking as factors in lung cancer. Amer. J. Public Health 44, N°2, 1954.

(8) BRESLOW L., BUECHLEY R. - Factors in coronary artery disease - Cigarette smoking and exercise. California Medicine 89, N°3 : 175-178, 1958.

+ (9) CASE R.A.M., LEA A.J. - Mustard gas poisoning, chronic bronchitis, and lung cancer. Brit. J. of Prev. & Soc. Med. 9, N°2, 1955.

(10) CLEMMESEN J. - Carcinoma of the breast : symposium results from statistical research. Brit. J. of Radiol. 21, N°252 : 583-590, 1948.

(11) CLEMMESEN J., LOCKWOOD K., NIELSEN A. - Smoking habits of patients with papilloma of urinary bladder. Danish Med. Bull. 5, N°3 : 123-128, 1958.

++ (12) COCHRAN W.G. - Matching in analytical studies. Amer J. of Public Health, Part I 43, N°6 : 684-691, 1953.

++ (13) COCHRAN W.G. - Some methods for strengthening the common X^2 tests Biometrics 10, N° 4 : 417-451, 1954.

(14) COHART E.M. - Socioeconomic distribution of stomach cancer in New-Haven. Cancer 7, N°3 : 455-461, 1954.

++ (15) CORNFIELD J. - A method of estimating comparative rates from clinical data. J. Nation. Canc. Instit. 2, N°6 : 1269-1275, 1951.

++ (16) CORNFIELD J., HAENSZEL W., HAMMOND E.C., LILIENFELD A.M., SHIMKIN M.B., WYNDER E.L. - Smoking and lung cancer : recent evidence and a discussion of some questions. J. of the Nat. Canc. Inst. 22, N°1 : 173-203, 1959

+ (17) COURT BROWN W.M., DOLL R. - Expectation of life and mortality from cancer among british radiologists. Brit. Med. J. ii : 181-187, 1958.

(18) DAWBER Th.R., MOORE F.E., MANN G.V. - Coronary heart disease in the Framingham study. Amer. H. of Public Health : 4-24, 1957.

(19) DENOIX P.F., SCHWARTZ D. - Tabac et cancer de la vessie. Bull. du Cancer 43, N°4 : 387-393, 1956.

+ (20) DENOIX P.F., SCHWARTZ D., ANGUERA G. - L'enquête

française sur l'étiologie du cancer broncho-pulmonaire - Analyse détaillée. Bull. du Cancer 45, N°1 : 1-37, 1958.

(21) DOLGOFF, SCHREK, BALLARD, BAKER - Tobacco smoking as an etiologic factor in disease - Coronary disease and hypertension. Angiology 3, N°4 : 323-334, 1952.

+ (22) DOLL R., HILL A.B. - A study of the aetiology of carcinoma of the lung. Brit. Med. J. 2 : 1271, 1952.

+ (23) DOLL R. - Mortality from lung cancer among abestos workers. Brit. J. Indust. Med. 12 : 81-86, 1955.

+ (24) DOLL R., HILL A.B. - Lung cancer and other causes of death in relation to smoking. Brit. Med. J. ii : 1071, 1956.

(25) DORN H.F. - Tobacco consumption and mortality from cancer and other diseases. Public Health Reports 74, N°7 : 581-593, 1959.

+ (26) DORN H.F. - Morbidity from cancer in the United States. Publ. Health Monogr. N°29, 1955.

++ (27) DORN H.F. - Some problems arising in prospective and retrospective studies of the etiology of disease. New England J. of Med. 261 : 571-579, 1959.

(28) ENGLISH, WILLIUS, BERKSON - Tobacco and coronary disease. J. Amer. Med. Assoc. 115, N°16, 1940.

+ (29) FRASER ROBERTS J.A. - Blood groups and susceptibility to disease. Brit. J. of Prevent & Soc. Med. 11, N°3, 1957.

(30) GILLIAM A.G. - Fertility and cancer of the breast and of uterine cervix - Comparisons between pregnancy rates among women with cancer at these and other sites. J. Nation. Canc. Inst. 12, N°2 : 287-304, 1951.

+ (31) HAENSZEL W., SHIMKIN M.B., MANTEL N. - A retrospective study of lung cancer in women. J. Nation. Canc. Inst. 21, N°5 : 825-842, 1958.

+ (32) HAMMOND E.C., HORN D. - Smoking and death rates - Report on fourty-four months of follow-up of 187 783 men. J. Amer. Med. Assoc. 166 : 1159-1308, 1958.

(33) HITCHCOCK C.R., SULLIVAN W.A., WANGENSTEEN O.H. - The value of achlorhydria as a screening test for gastric cancer. Gastroenterology 29, N°4 : 621-628, 1955.

(34) JONES E.G., MACDONALD I., BRESLOW L. - Study of epidemiologic factors in carcinoma of uterine cervix. Amer. J. Obst. & Gynec. 76 : 1-10, 1958.

+ (35) KORTEWEG R. - The significance of selection in prospective investigations into an association between smoking and lung cancer. Brit. J. Cancer 10 : 282-291, 1956.

(36) LAMY M., SEROR M.E. - Les embryopathies d'origine rubéolique. Semaine Hôpit. Paris, N°36, 1956.

(37) LANE-CLAYPON J.E. - A further report on cancer of the breast, with special referance to its associated antecedent conditions. Rept. Publ. Health & M. Subj., N°32 : 1-189, 1926.

+ (38) LEDERMANN - Cancers - Tabac - Vin & Alcool. Concours Médical, N°11 & 12, 1955.

+ (39) LEVIN M.L. - The occurence of lung cancer in man. Acta Intern. Union against Cancer 9, N°3, 1953.

(40) LILIENFELD A., LEVIN M.L., MOORE G.E. - The association of smoking with cancer of the urinary bladder in humans Arch. of Intern. Med. 98 : 129-135, 1956.

+ (41) LILIENFELD A.M. - Emotional and other selected characteristics of cigarette smokers and nonsmokers as related to epidemiological studies of lung cancer and other diseases. J. Nation. Canc. Inst. 22, N°2 : 259-282, 1959.

++ (42) LOMBARD H.L., DOERING C.R. - Treatment of the fourfold table by partial association and partial correlation as it relates to public health problems. Biometrics 3, N°3 : 123-128, 1947.

(43) LOMBARD, POTTER - Epidemiological aspects of cancer of cervix. II. Hereditary and environmental factors. Cancer 3 : 960-968, 1950.

(44) LOWE C.R. - An association between smoking and respiratory tuberculosis. Brit. Med. J. : 1082-1086, 1956.

+ (45) MACHT S.H., LAWRENCE P.S. - National survey of congenital malformations resulting from exposure to roentgen radiation. Amer. J. Roent. & Radiumtherapy 73, N°3, 1955.

+ (46) MACKLIN M.T. - Comparison of the number of breast-cancer deaths observed in relatives of breast-cancer patients, and the number expected on the basis of mortality rates. J. Canc. Inst. 22, N°5 : 927-951, 1959.

++ (47) MAINLAND D. - The risk of fallacious conclusions from autopsy data on the incidence of diseases with applications to heart disease. Amer Heart J. 45, N°5 : 644-654, 1953.

(48) MALIPHANT R.G. - The incidence of the cancer of the uterine

cervix. Brit. Med. J., N°4613 : 978-981, 1949.

++ (49) MANTEL N., HAENSZEL W. - Statistical aspects of the analysis of data from retrospective studies of disease. J. Nation. Canc. Inst. 22, N°4 : 719-748, 1959.

(50) MILLS C.A., MILLS PORTER - Tobacco smoking habits and cancer of the mouth and respiratory system. Cancer Research 10, N°9 : 539-542, 1950.

(51) MOSBECH J., VIDEBAEK A. - Mortality from and risk of carcinoma among patients with pernicious anaemia. Brit. Med. J. 2 : 390, 1950.

(52) MUSTACCHI - Cancer of the bladder and infestation with "schistosoma hematobium". J. Nation. Canc. Inst. 20 : 825-842, 1958.

++ (53) NEYMAN J. - Statistics - Servant of all sciences. Science 122 : 401-406, 1955.

(54) PEQUIGNOT G. - Enquête par interrogatoire sur les circonstances diététiques de la cirrhose alcoolique en France. Bull. Inst. Nation. Hygiène 13 : 719-739, 1958.

+ (55) SADOWSKY D.A., CORNFIELD J., GILLIAM A.G. - The statistical association between smoking and carcinoma of the lung. J. Nation. Canc. Inst. 13 : 1237-1258, 1953.

+ (56) SANGHVI L.D., RAO K.C.M., KHANOLKAR V.R. - Smoking and chewing of tobacco in relation to cancer of the upper alimentary tract. Brit. Med. J. i, N°4922 : 1111-1114, 1955.

++ (57) SCHWARTZ D., ANGUERA G. - Une cause de biais dans certaines enquêtes médicales : le temps de séjour des malades à l'hôpital. Comm. Inst. Intern. Statist. 30e Session - Stockholm 1957.

(58) SCHWARTZ D., DENOIX P.F., ANGUERA G. - Recherches des localisations du cancer associées aux facteurs tabac et alcool chez l'homme. Bull. du Cancer 44, N°2 : 336-361, 1957.

(59) SCHWARTZ D., DENOIX P.F., ROUQUETTE C. - Enquête sur l'étiologie des cancers génitaux de la femme. Bull. du Cancer 45, N°4 : 476-493, 1958.

(60) SCHWARTZ D., DENOIX P.F. - L'enquête française sur l'étiologie du cancer broncho-pulmonaire - Rôle du tabac. Semaine Hôpit. Paris, N°62/7 : 424-437, 1957.

++ (61) SHEPS M.C. - An examination of some methods of comparing several rates or proportions. Biometrics 15, N°1 : 87-97, 1959.

(62) STEWART A., WEBB J., Coll. - Malignant disease in childhood and diagnostic irradiation in utero. Lancet 2 : 447, 1956.

(63) STOCKS P. - The epidemiology of carcinoma of the breast. The Practitioner 179 : 233-240, 1957.

(64) VIDEBAEK A., MOSBECH J. - Genetic causal factors in cancer of the stomach. Danish Med. Bull. 1, N°7 : 189-193, 1954.

(65) WHITE, SHARBER - Tabac, alcool et angine de poitrine. J. Amer. Med. Associat. N°102 : 655, 1934.

(66) WHITE C., BAILAR III J.C. - Rétrospective and prospective methods of studying association in medicine. Amer J. Public Health & Nat. Health 46, N°1 : 35-44, 1956.

(67) WYNDER E.L., GRAHAM E.A. - Tobacco smoking as a possible etiologic factor in bronchiogenic carcinoma. J. Amer. Med. Assoc. 143, N°4 : 329-336, 1950.

(68) WYNDER E.L., CORNFIELD J., SHROFF P. - A study of environmental factors in carcinoma of the cervix. Amer. J. Obst. & Gynec. St-Louis 68, N°4 : 1016-1052, 1954.

(69) WYNDER E.L., BROSS I.J., DAY E. - Epidemiological approach to the etiology of cancer of the larynx. J. Amer. Med. Assoc. 160 : 1384-1391, 1956.

(70) WYNDER E.L., BROSS I.J., CORNFIELD J., O'DONNELL - Lung cancer in woman - A study of environmental factors. New England J. Medic. 225 : 1111-1121, 1956.

(71) WYNDER E.L., BROSS I.J., FELDMAN R.M. - A study of the etiological factors in cancer of the mouth. Cancer 10, N°6 : 1300-1323, 1957.

IMP. LOUIS-JEAN — GAP
Depôt légal n° 138 — 1960